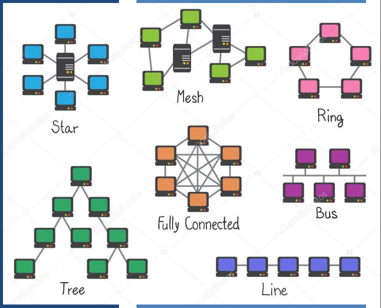
TCSS 558: APPLIED DISTRIBUTED COMPUTING

Ch. 3 - Processes: Servers

Ch. 4 - Communication

Wes J. Lloyd
School of Engineering & Technology (SET)
University of Washington - Tacoma

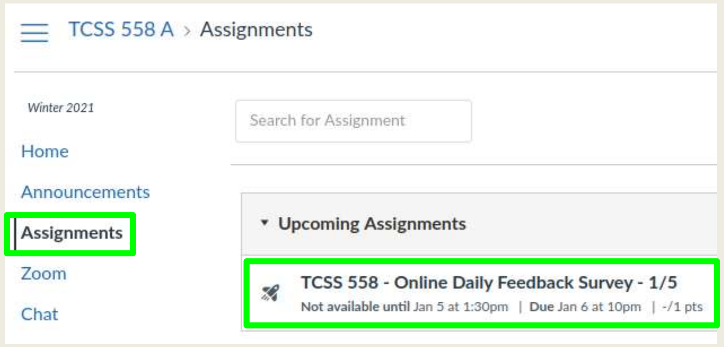


OBJECTIVES – 2/18

- Questions from 2/16
- Verify Midterm Scoring
- Assignment 1: Key/Value Store
  - Java Maven project template files posted
- Chapter 3: Processes
  - Chapter 3.4: Servers
  - Chapter 3.5: Resource (Code) Migration (*light-review)*
- Chapter 4: Communication
  - Chapter 4.1: Foundations
  - Chapter 4.2: RPC
  - Chapter 4.3: Message Oriented Communication

ONLINE DAILY FEEDBACK SURVEY

- Daily Feedback Quiz in Canvas – Available After Each Class
- Extra credit available for completing surveys ***ON TIME***
- Tuesday surveys: due by ~ Wed @ 10p
- Thursday surveys: due ~ Mon @ 10p


TCSS 558 A > Assignments

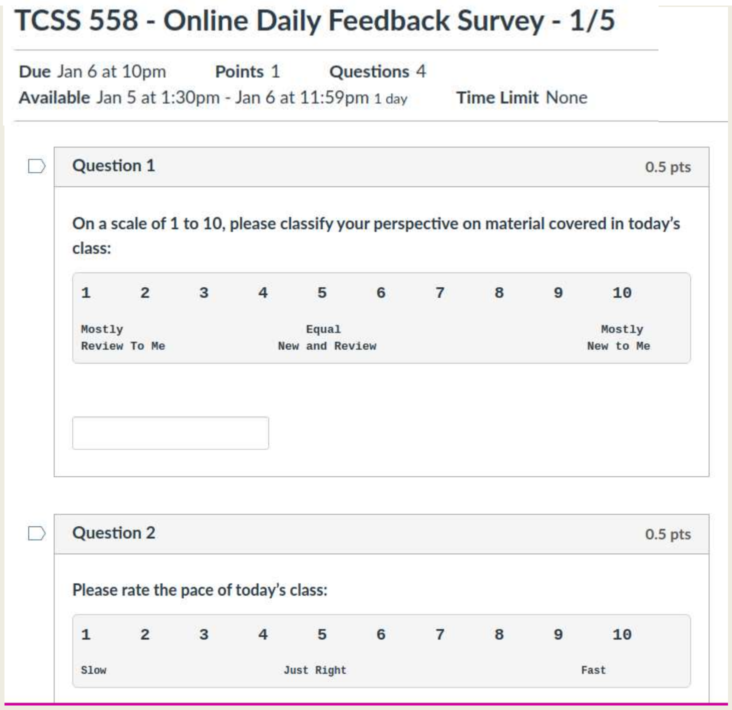
Winter 2021

Home
Announcements
Assignments
Zoom
Chat

Search for Assignment

Upcoming Assignments

TCSS 558 - Online Daily Feedback Survey - 1/5
Not available until Jan 5 at 1:30pm | Due Jan 6 at 10pm | -/1 pts


TCSS 558 - Online Daily Feedback Survey - 1/5

Due Jan 6 at 10pm **Points** 1 **Questions** 4
Available Jan 5 at 1:30pm - Jan 6 at 11:59pm 1 day **Time Limit** None

Question 1 0.5 pts

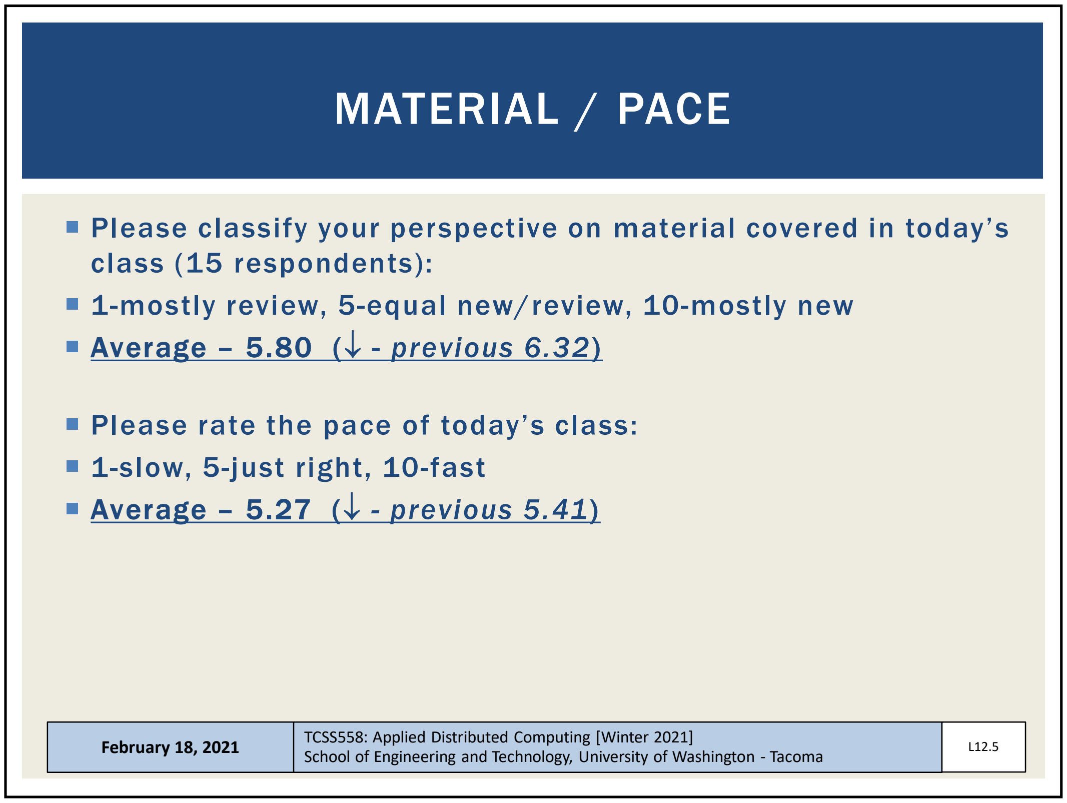
On a scale of 1 to 10, please classify your perspective on material covered in today's class:

```
1     2     3     4     5     6     7     8     9     10

Mostly                  Equal                     Mostly
Review To Me       New and Review                New to Me
```

Question 2 0.5 pts

Please rate the pace of today's class:

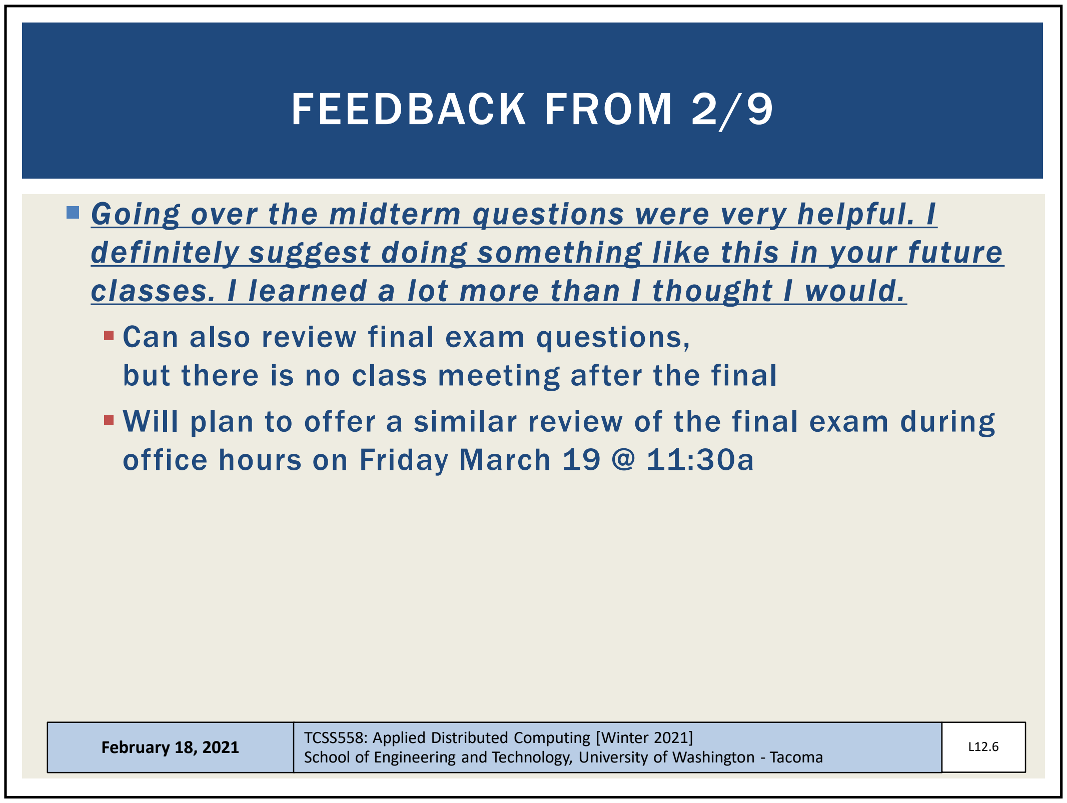
```
1     2     3     4     5     6     7     8     9     10

Slow                 Just Right                  Fast
```



MATERIAL / PACE

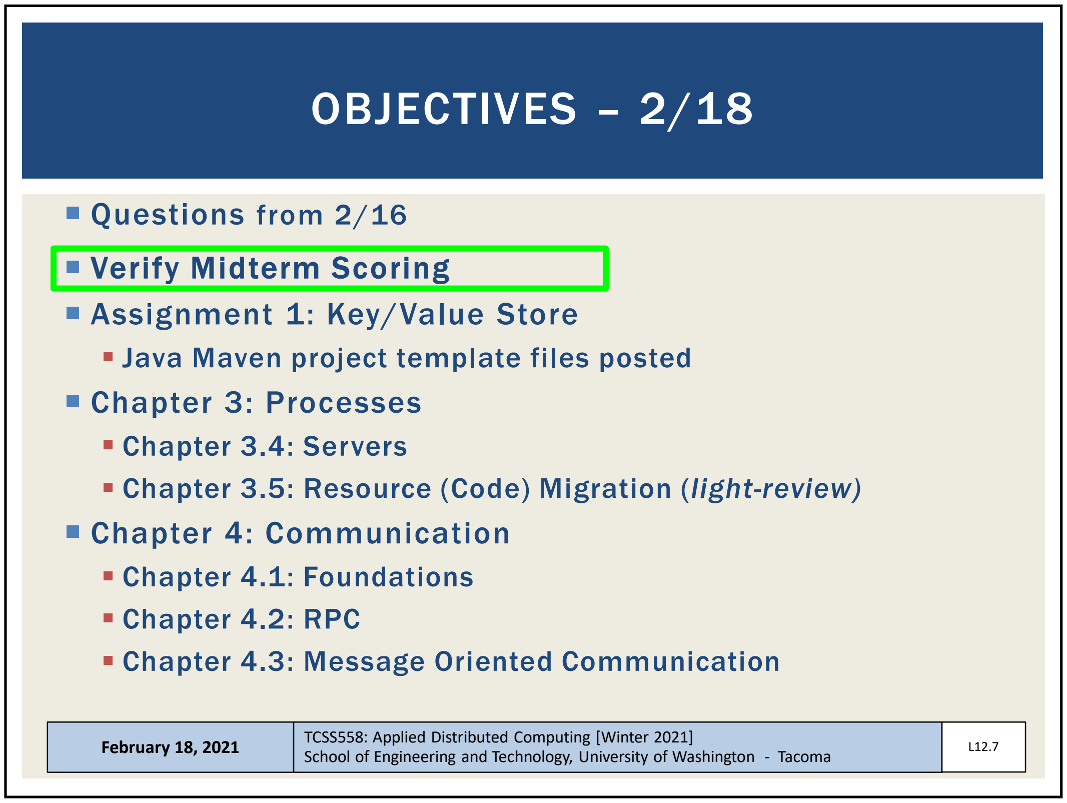
- Please classify your perspective on material covered in today's class (15 respondents):
- 1-mostly review, 5-equal new/review, 10-mostly new
- **Average – 5.80 (↓ - *previous 6.32*)**

- Please rate the pace of today's class:
- 1-slow, 5-just right, 10-fast
- **Average – 5.27 (↓ - *previous 5.41*)**



FEEDBACK FROM 2/9

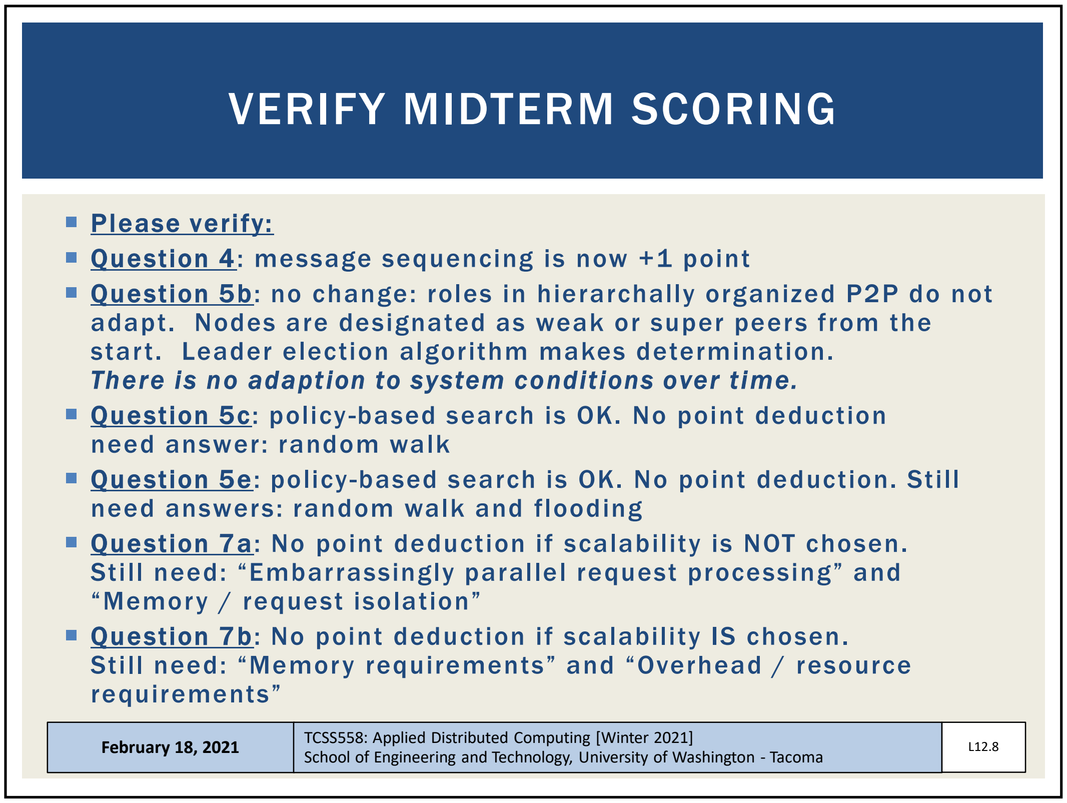
- ***Going over the midterm questions were very helpful. I definitely suggest doing something like this in your future classes. I learned a lot more than I thought I would.***
  - Can also review final exam questions, but there is no class meeting after the final
  - Will plan to offer a similar review of the final exam during office hours on Friday March 19 @ 11:30a



OBJECTIVES – 2/18

- Questions from 2/16
- Verify Midterm Scoring
- Assignment 1: Key/Value Store
  - Java Maven project template files posted
- Chapter 3: Processes
  - Chapter 3.4: Servers
  - Chapter 3.5: Resource (Code) Migration (*light-review)*
- Chapter 4: Communication
  - Chapter 4.1: Foundations
  - Chapter 4.2: RPC
  - Chapter 4.3: Message Oriented Communication

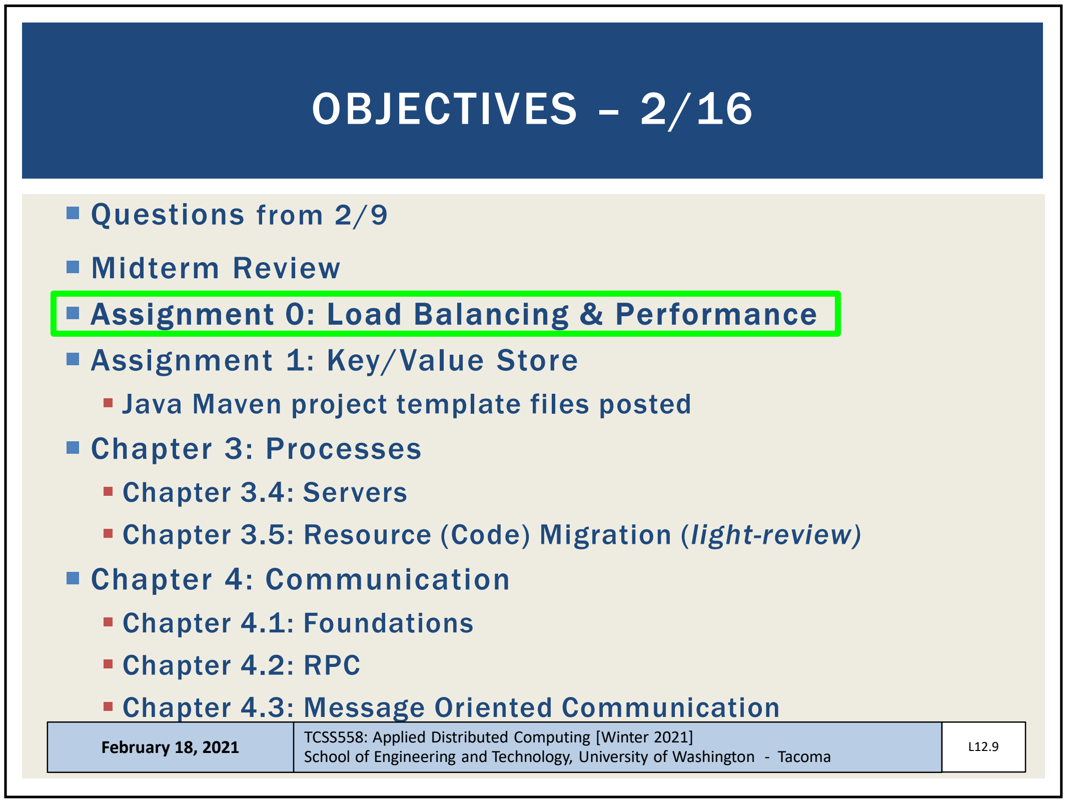
February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.7



VERIFY MIDTERM SCORING

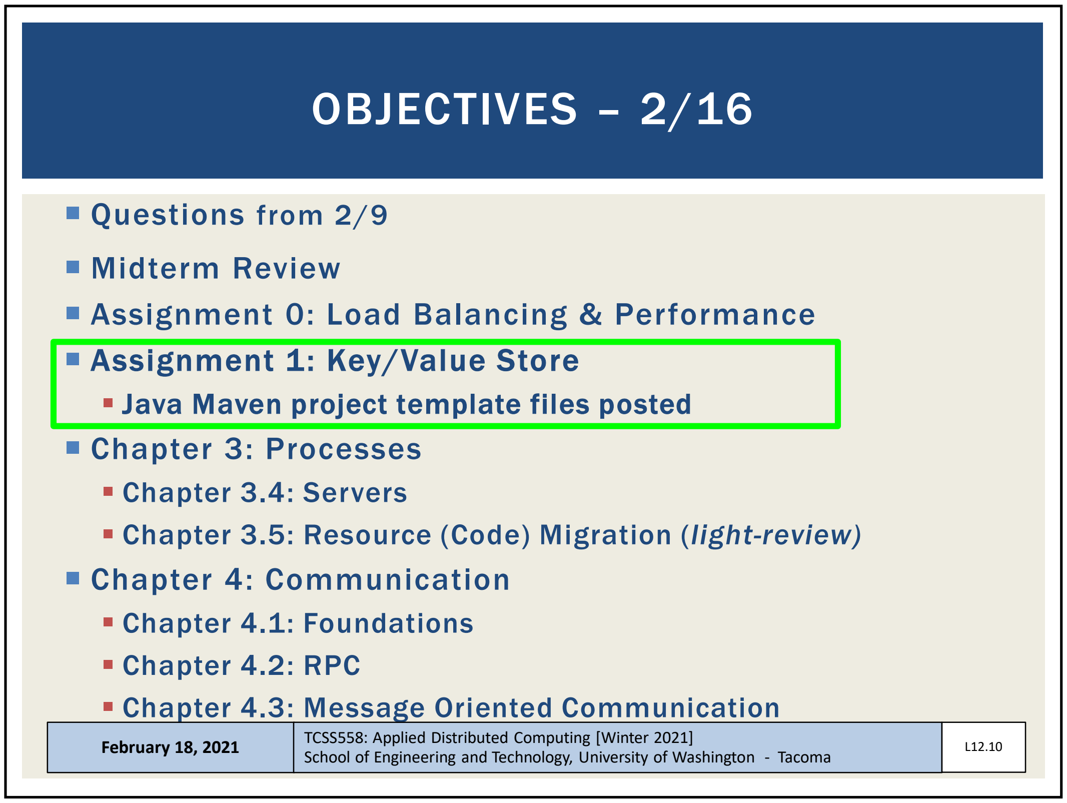
- Please verify:
- Question 4: message sequencing is now +1 point
- Question 5b: no change: roles in hierarchally organized P2P do not adapt. Nodes are designated as weak or super peers from the start. Leader election algorithm makes determination.
  There is no adaption to system conditions over time.
- Question 5c: policy-based search is OK. No point deduction need answer: random walk
- Question 5e: policy-based search is OK. No point deduction. Still need answers: random walk and flooding
- Question 7a: No point deduction if scalability is NOT chosen. Still need: “Embarrassingly parallel request processing” and “Memory / request isolation”
- Question 7b: No point deduction if scalability IS chosen. Still need: “Memory requirements” and “Overhead / resource requirements”

February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.8



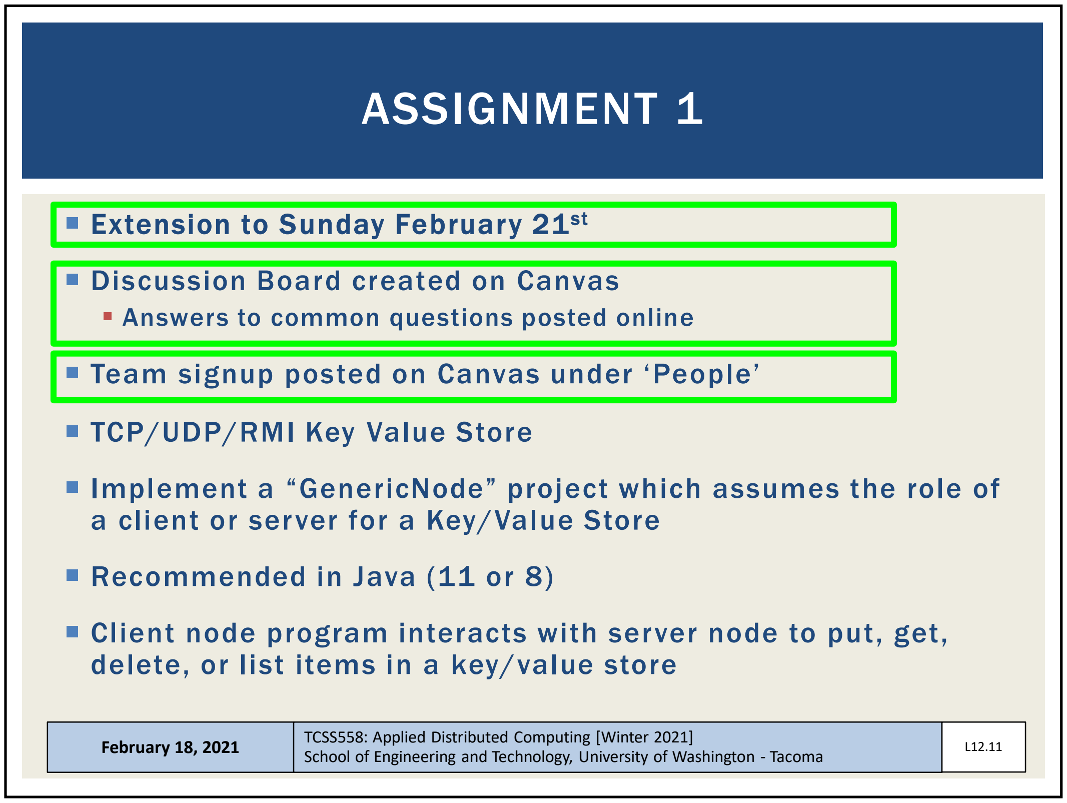
OBJECTIVES – 2/16

- Questions from 2/9
- Midterm Review
- **Assignment 0: Load Balancing & Performance**
- Assignment 1: Key/Value Store
  - Java Maven project template files posted
- Chapter 3: Processes
  - Chapter 3.4: Servers
  - Chapter 3.5: Resource (Code) Migration (*light-review)*
- Chapter 4: Communication
  - Chapter 4.1: Foundations
  - Chapter 4.2: RPC
  - Chapter 4.3: Message Oriented Communication



OBJECTIVES – 2/16

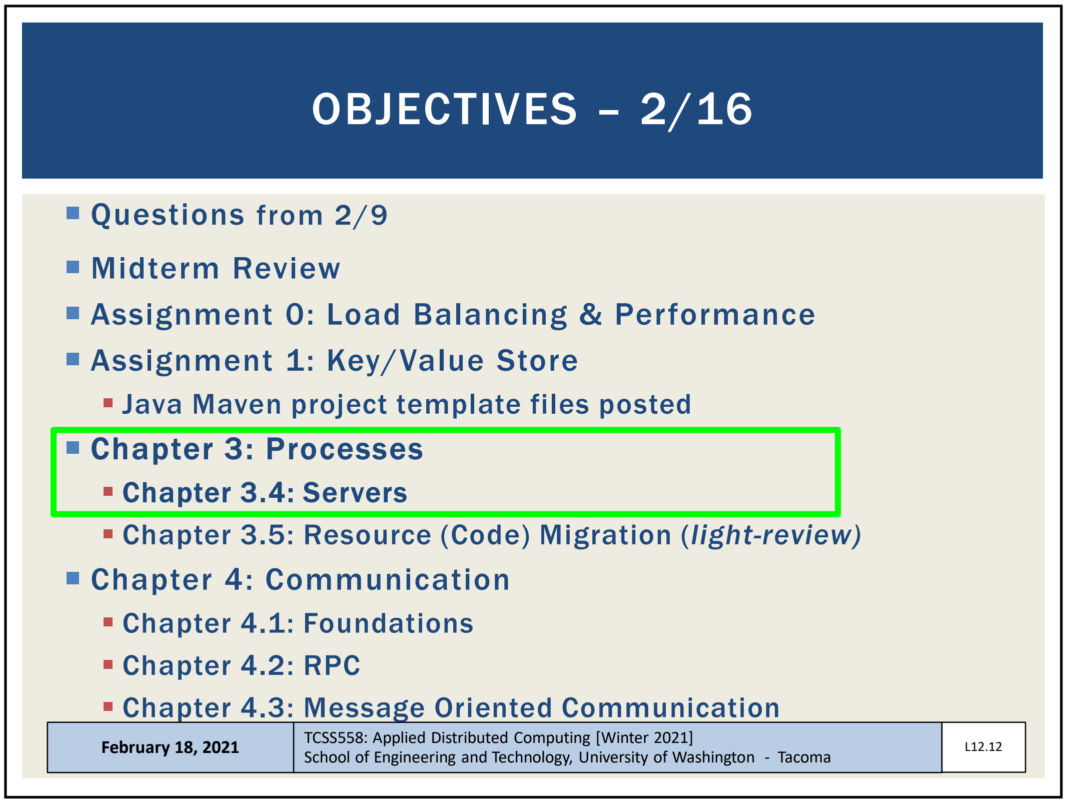
- Questions from 2/9
- Midterm Review
- Assignment 0: Load Balancing & Performance
- **Assignment 1: Key/Value Store**
  - **Java Maven project template files posted**
- Chapter 3: Processes
  - Chapter 3.4: Servers
  - Chapter 3.5: Resource (Code) Migration (*light-review)*
- Chapter 4: Communication
  - Chapter 4.1: Foundations
  - Chapter 4.2: RPC
  - Chapter 4.3: Message Oriented Communication



ASSIGNMENT 1

- Extension to Sunday February 21st
- Discussion Board created on Canvas
  - Answers to common questions posted online
- Team signup posted on Canvas under 'People'
- TCP/UDP/RMI Key Value Store
- Implement a "GenericNode" project which assumes the role of a client or server for a Key/Value Store
- Recommended in Java (11 or 8)
- Client node program interacts with server node to put, get, delete, or list items in a key/value store

February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.11



OBJECTIVES – 2/16

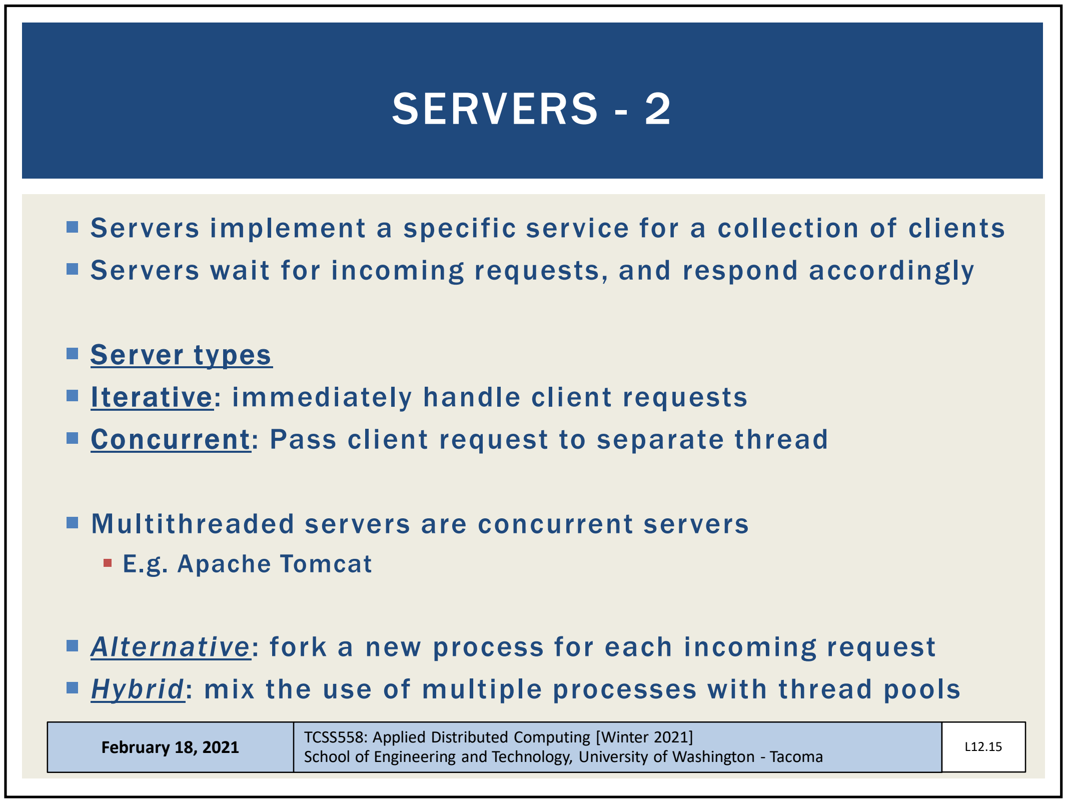
- Questions from 2/9
- Midterm Review
- Assignment 0: Load Balancing & Performance
- Assignment 1: Key/Value Store
  - Java Maven project template files posted
- Chapter 3: Processes
  - Chapter 3.4: Servers
  - Chapter 3.5: Resource (Code) Migration (*light-review)*
- Chapter 4: Communication
  - Chapter 4.1: Foundations
  - Chapter 4.2: RPC
  - Chapter 4.3: Message Oriented Communication

February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.12

CH. 3.4: SERVERS

SERVERS

- Cloud & Distributed Systems – rely on **Linux**
- http://www.zdnet.com/article/it-runs-on-the-cloud-and-the-cloud-runs-on-linux-any-questions/
- IT is moving to the cloud. And, what powers the cloud?
  - **Linux**
- Uptime Institute survey - 1,000 IT executives (2016)
  - 50% of IT executives – plan to migrate majority of IT workloads to off-premise to cloud or colocation sites
  - 23% expect the shift in 2017, 70% by 2020…
- Docker on Windows / Mac OS X
  - Based on **Linux**
  - Mac: Hyperkit Linux VM
  - Windows: Hyper-V Linux VM



SERVERS - 2

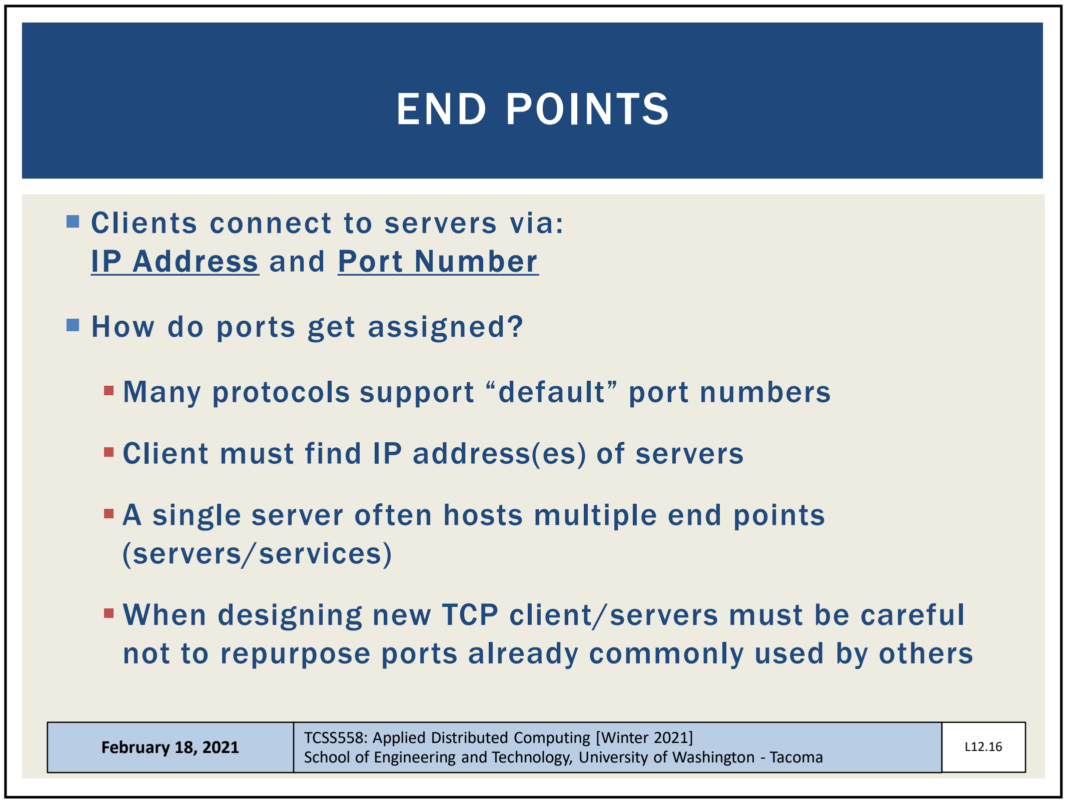
- Servers implement a specific service for a collection of clients
- Servers wait for incoming requests, and respond accordingly

- **Server types**
- **Iterative**: immediately handle client requests
- **Concurrent**: Pass client request to separate thread

- Multithreaded servers are concurrent servers
  - E.g. Apache Tomcat

- ***Alternative***: fork a new process for each incoming request
- ***Hybrid***: mix the use of multiple processes with thread pools

February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.15

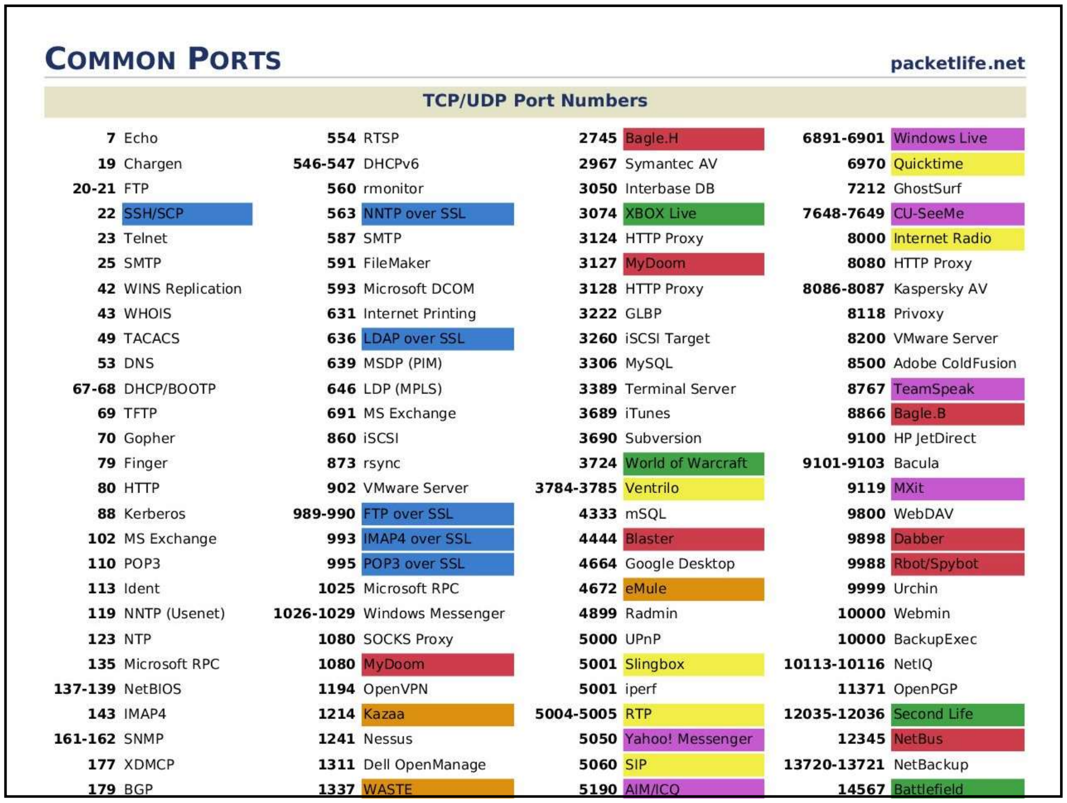


END POINTS

- Clients connect to servers via:
  IP Address and **Port Number**

- How do ports get assigned?
  - Many protocols support "default" port numbers
  - Client must find IP address(es) of servers
  - A single server often hosts multiple end points (servers/services)
  - When designing new TCP client/servers must be careful not to repurpose ports already commonly used by others

February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.16

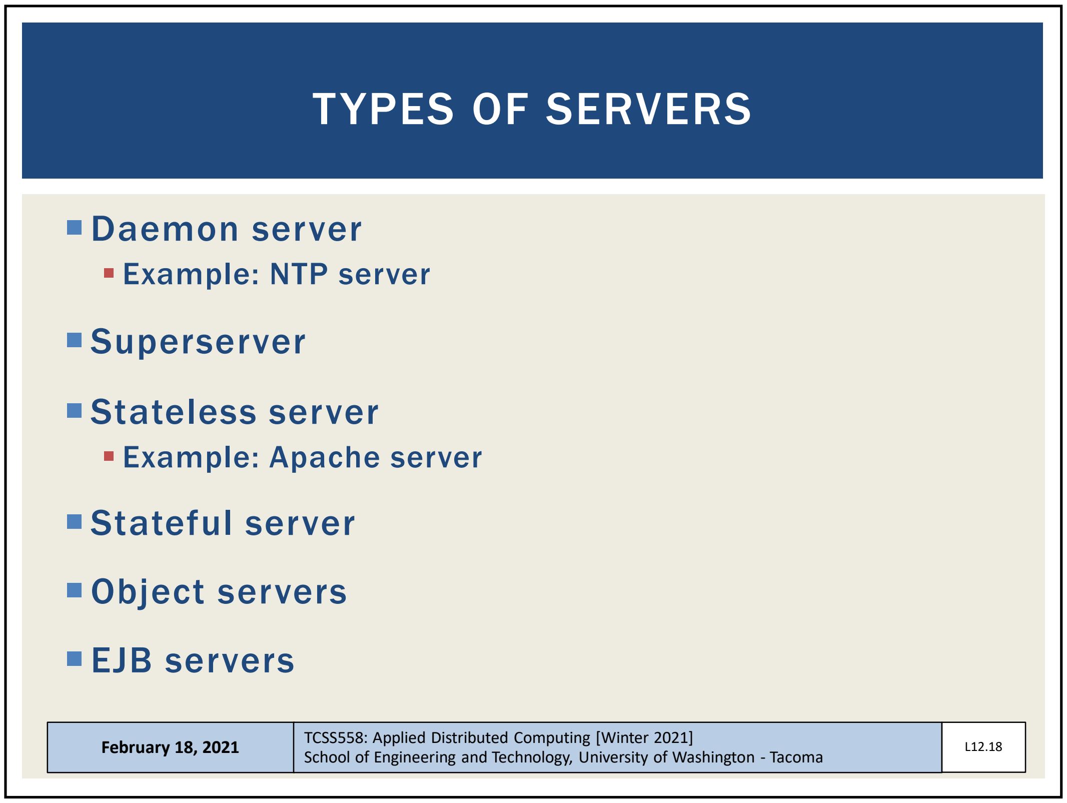


Common Ports

packetlife.net

TCP/UDP Port Numbers

7	Echo	554	RTSP	2745	Bagle.H	6891-6901	Windows Live
19	Chargen	546-547	DHCPv6	2967	Symantec AV	6970	Quicktime
20-21	FTP	560	rmonitor	3050	Interbase DB	7212	GhostSurf
22	SSH/SCP	563	NNTP over SSL	3074	XBOX Live	7648-7649	CU-SeeMe
23	Telnet	587	SMTP	3124	HTTP Proxy	8000	Internet Radio
25	SMTP	591	FileMaker	3127	MyDoom	8080	HTTP Proxy
42	WINS Replication	593	Microsoft DCOM	3128	HTTP Proxy	8086-8087	Kaspersky AV
43	WHOIS	631	Internet Printing	3222	GLBP	8118	Privoxy
49	TACACS	636	LDAP over SSL	3260	iSCSI Target	8200	VMware Server
53	DNS	639	MSDP (PIM)	3306	MySQL	8500	Adobe ColdFusion
67-68	DHCP/BOOTP	646	LDP (MPLS)	3389	Terminal Server	8767	TeamSpeak
69	TFTP	691	MS Exchange	3689	iTunes	8866	Bagle.B
70	Gopher	860	iSCSI	3690	Subversion	9100	HP JetDirect
79	Finger	873	rsync	3724	World of Warcraft	9101-9103	Bacula
80	HTTP	902	VMware Server	3784-3785	Ventrilo	9119	MXit
88	Kerberos	989-990	FTP over SSL	4333	mSQL	9800	WebDAV
102	MS Exchange	993	IMAP4 over SSL	4444	Blaster	9898	Dabber
110	POP3	995	POP3 over SSL	4664	Google Desktop	9988	Rbot/Spybot
113	Ident	1025	Microsoft RPC	4672	eMule	9999	Urchin
119	NNTP (Usenet)	1026-1029	Windows Messenger	4899	Radmin	10000	Webmin
123	NTP	1080	SOCKS Proxy	5000	UPnP	10000	BackupExec
135	Microsoft RPC	1080	MyDoom	5001	Slingbox	10113-10116	NetIQ
137-139	NetBIOS	1194	OpenVPN	5001	iperf	11371	OpenPGP
143	IMAP4	1214	Kazaa	5004-5005	RTP	12035-12036	Second Life
161-162	SNMP	1241	Nessus	5050	Yahoo! Messenger	12345	NetBus
177	XDMCP	1311	Dell OpenManage	5060	SIP	13720-13721	NetBackup
179	BGP	1337	WASTE	5190	AIM/ICQ	14567	Battlefield



TYPES OF SERVERS

- Daemon server
  - Example: NTP server
- Superserver
- Stateless server
  - Example: Apache server
- Stateful server
- Object servers
- EJB servers

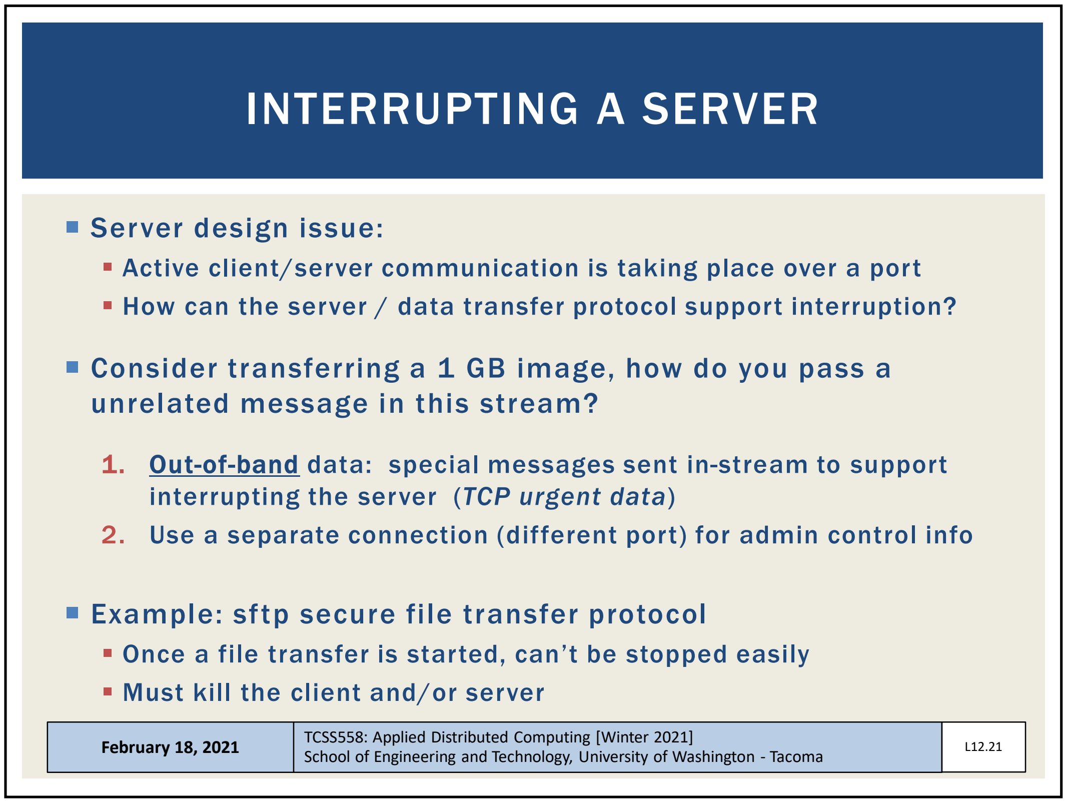
February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.18

NTP EXAMPLE

- Daemon servers
  - Run locally on Linux
  - Track current server end points (outside servers)
  - Example: network time protocol (ntp) daemon
    - Listen locally on specific port (ntp is 123)
    - Daemons routes local client traffic to the configured endpoint servers
    - University of Washington: time.u.washington.edu
    - Example "`ntpq -p`"
      - Queries local ntp daemon, routes traffic to configured server(s)

SUPERSERVER

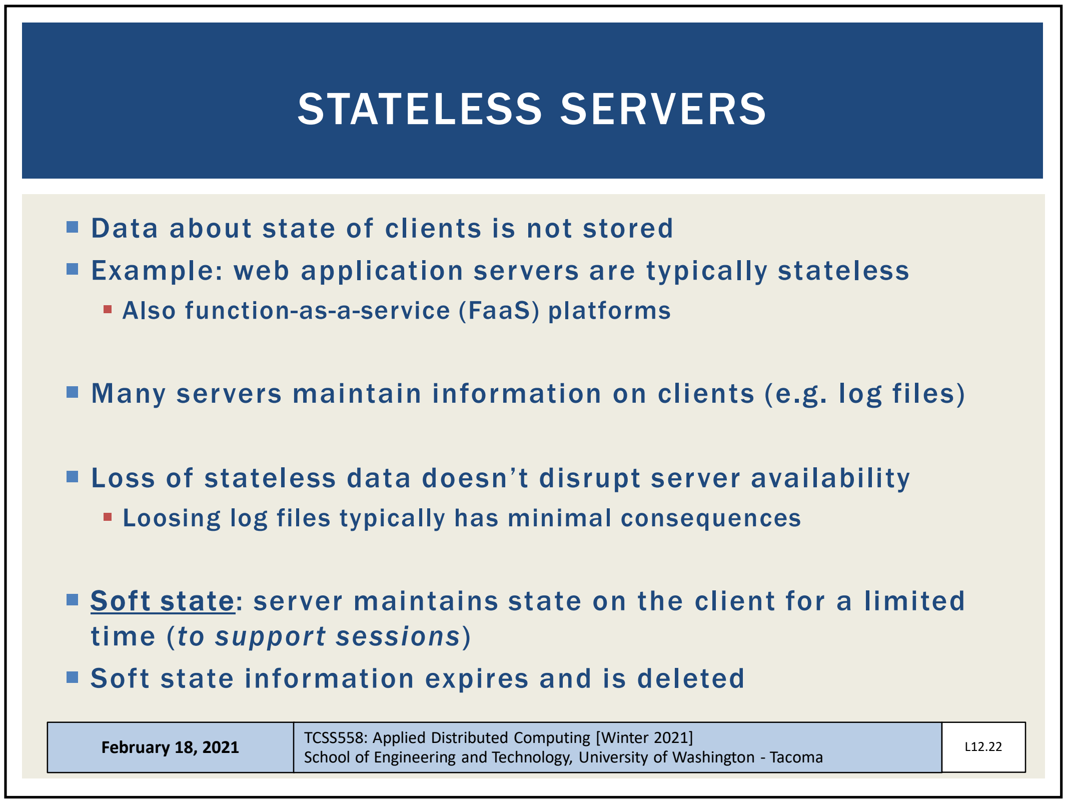
- Linux inetd / xinetd
  - Single superserver
  - Extended internet service daemon
  - Not installed by default on Ubuntu
  - Intended for use on server machines
  - Used to configure box as a server for multiple internet services
    - E.g. ftp, pop, telnet
  - inetd daemon responds to multiple endpoints for multiple services
  - Requests fork a process to run required executable program
- Check what ports you're listening on:
  - `sudo netstat -tap | grep LISTEN`



INTERRUPTING A SERVER

- Server design issue:
  - Active client/server communication is taking place over a port
  - How can the server / data transfer protocol support interruption?
- Consider transferring a 1 GB image, how do you pass a unrelated message in this stream?
  1. <u>Out-of-band</u> data: special messages sent in-stream to support interrupting the server (*TCP urgent data*)
  2. Use a separate connection (different port) for admin control info
- Example: sftp secure file transfer protocol
  - Once a file transfer is started, can't be stopped easily
  - Must kill the client and/or server

February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.21



STATELESS SERVERS

- Data about state of clients is not stored
- Example: web application servers are typically stateless
  - Also function-as-a-service (FaaS) platforms
- Many servers maintain information on clients (e.g. log files)
- Loss of stateless data doesn't disrupt server availability
  - Loosing log files typically has minimal consequences
- <u>Soft state</u>: server maintains state on the client for a limited time (*to support sessions*)
- Soft state information expires and is deleted

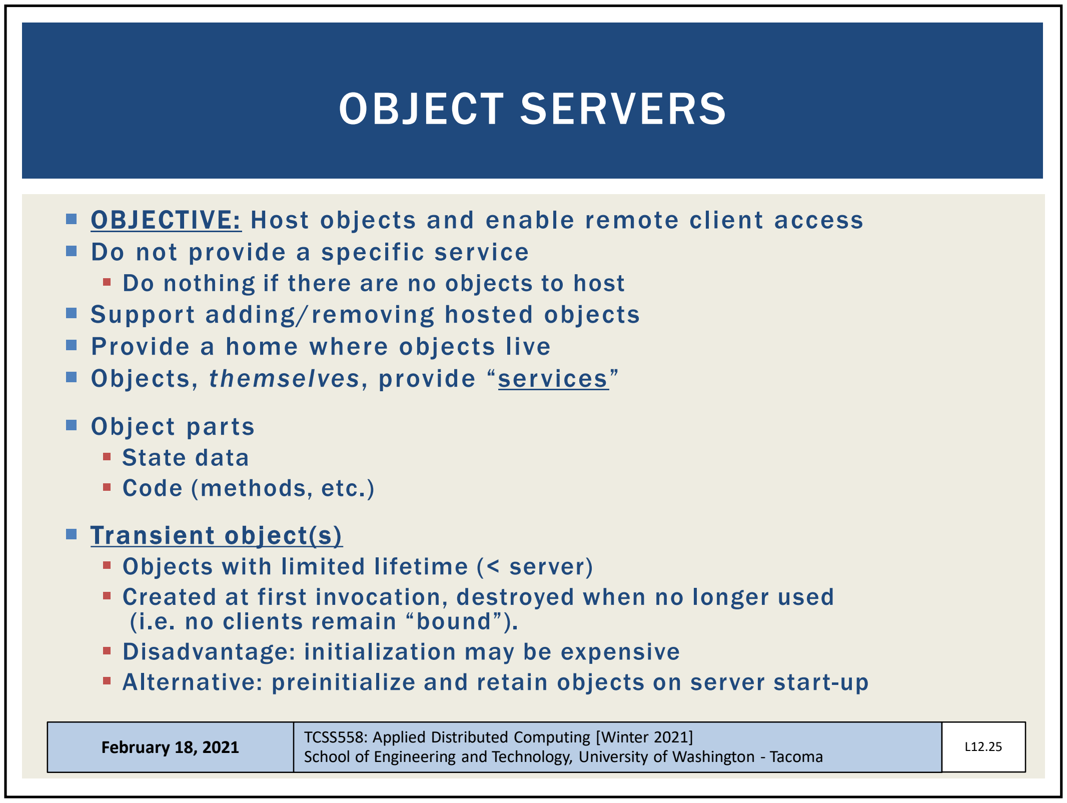
February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.22

STATEFUL SERVERS

- Maintain persistent information about clients
- Information must be explicitly deleted by the server
- Example:
  File server - allows clients to keep local file copies for RW
- Server tracks client file permissions and most recent versions
  - Table of (client, file) entries
- If server crashes data must be recovered
- Entire state before a crash must be restored
- Fault tolerance - *Ch. 8*

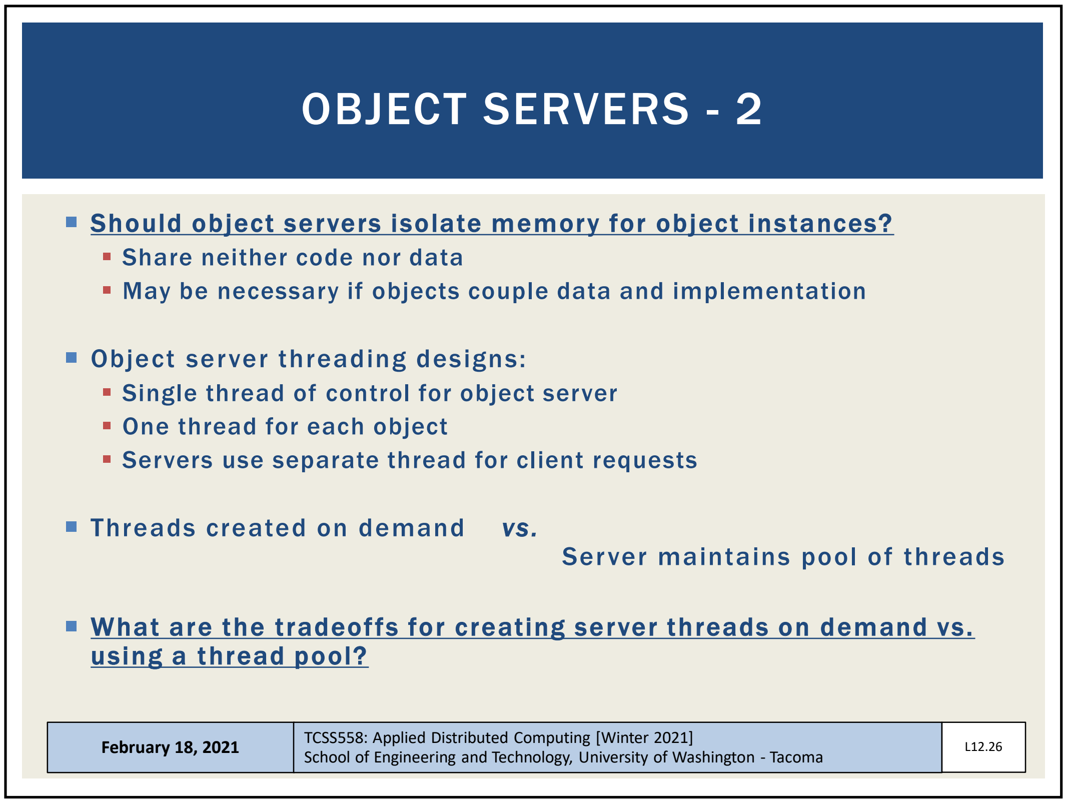
STATEFUL SERVERS - 2

- Session state
  - Tracks series of operations by a single user
  - Maintained temporarily, not indefinitely
  - Often retained for multi-tier client server applications
  - Minimal consequence if session state is lost
  - Clients must start over, reinitialize sessions
- Permanent state
  - Customer information, software keys
- Client-side cookies
  - When servers don’t maintain client state, clients can store state locally in “cookies”
  - Cookies are not executable, simply client-side data



OBJECT SERVERS

- **OBJECTIVE:** Host objects and enable remote client access
- Do not provide a specific service
  - Do nothing if there are no objects to host
- Support adding/removing hosted objects
- Provide a home where objects live
- Objects, *themselves*, provide "services"
- Object parts
  - State data
  - Code (methods, etc.)
- Transient object(s)
  - Objects with limited lifetime (< server)
  - Created at first invocation, destroyed when no longer used (i.e. no clients remain "bound").
  - Disadvantage: initialization may be expensive
  - Alternative: preinitialize and retain objects on server start-up



OBJECT SERVERS - 2

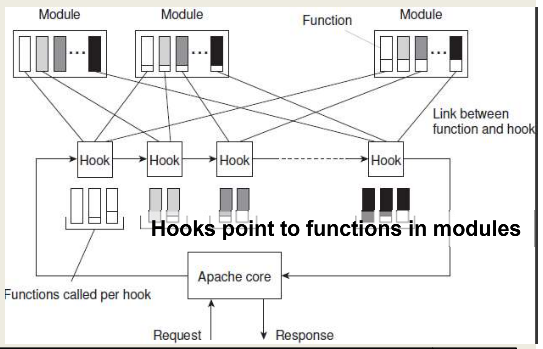
- **Should object servers isolate memory for object instances?**
  - Share neither code nor data
  - May be necessary if objects couple data and implementation
- Object server threading designs:
  - Single thread of control for object server
  - One thread for each object
  - Servers use separate thread for client requests
- Threads created on demand *vs.* Server maintains pool of threads
- **What are the tradeoffs for creating server threads on demand vs. using a thread pool?**

EJB – ENTERPRISE JAVA BEANS

- EJB- specialized Java object hosted by a EJB web container
- 4 types: stateless, stateful, entity, and message-driven beans
- Provides "middleware" standard (framework) for implementing back-ends of enterprise applications
- EJB web application containers integrate support for:
  - Transaction processing
  - Persistence
  - Concurrency
  - Event-driven programming
  - Asynchronous method invocation
  - Job scheduling
  - Naming and discovery services (JNDI)
  - Interprocess communication
  - Security
  - Software component deployment to an application server

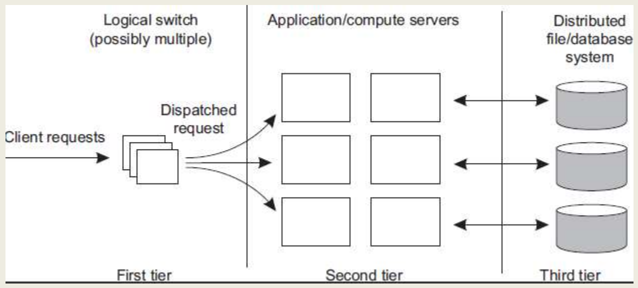
APACHE WEB SERVER

- Highly configurable, extensible, platform independent
- Supports TCP HTTP protocol communication
- Uses hooks – placeholders for group of functions
- Requests processed in phases by hooks
- Many hooks:
  - Translate a URL
  - Write info to log
  - Check client ID
  - Check access rights
- Hooks processed in order enforcing flow-of-control
- Functions in replaceable modules


Module
Module
Function
Module
Link between function and hook
Hook
Hook
Hook
Hook
Hooks point to functions in modules
Apache core
Functions called per hook
Request
Response

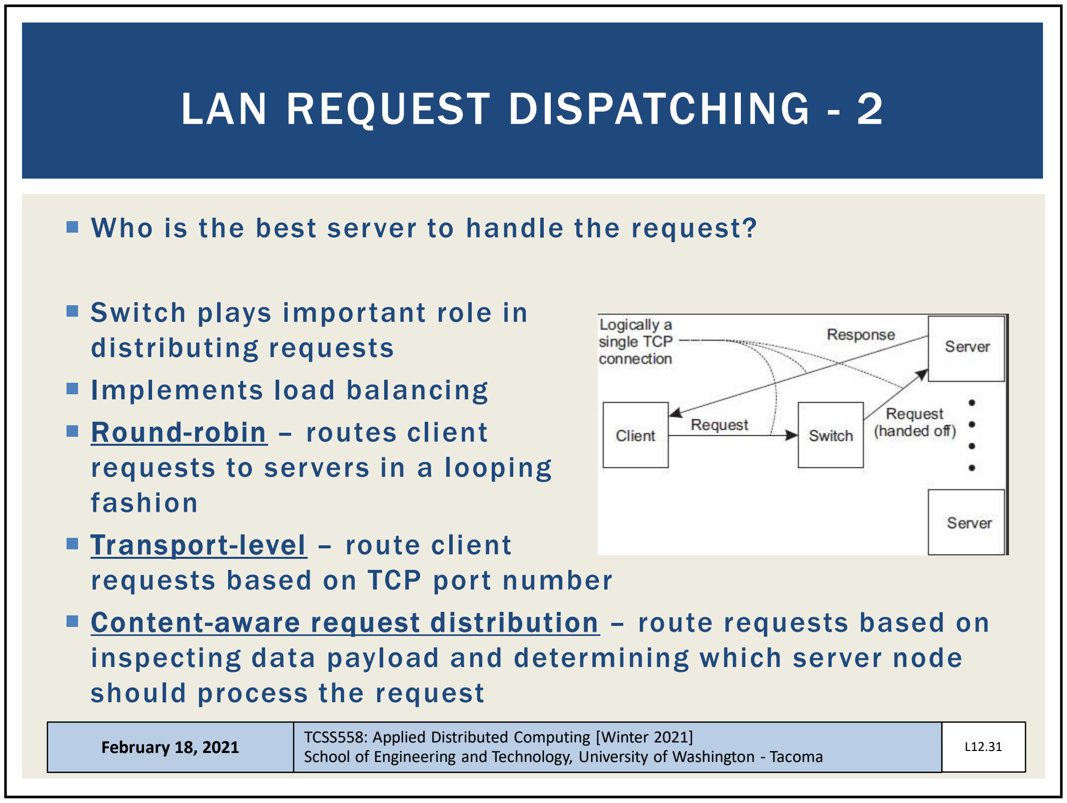
SERVER CLUSTERS

- Hosted across an LAN or WAN
- Collection of interconnected machines
- Can be organized in tiers:
  - Web server → app server → DB server
  - App and DB server sometimes integrated



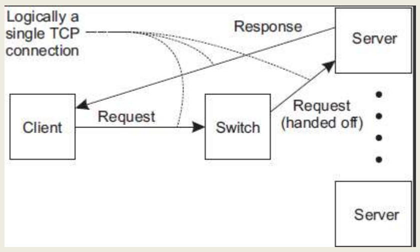
LAN REQUEST DISPATCHING

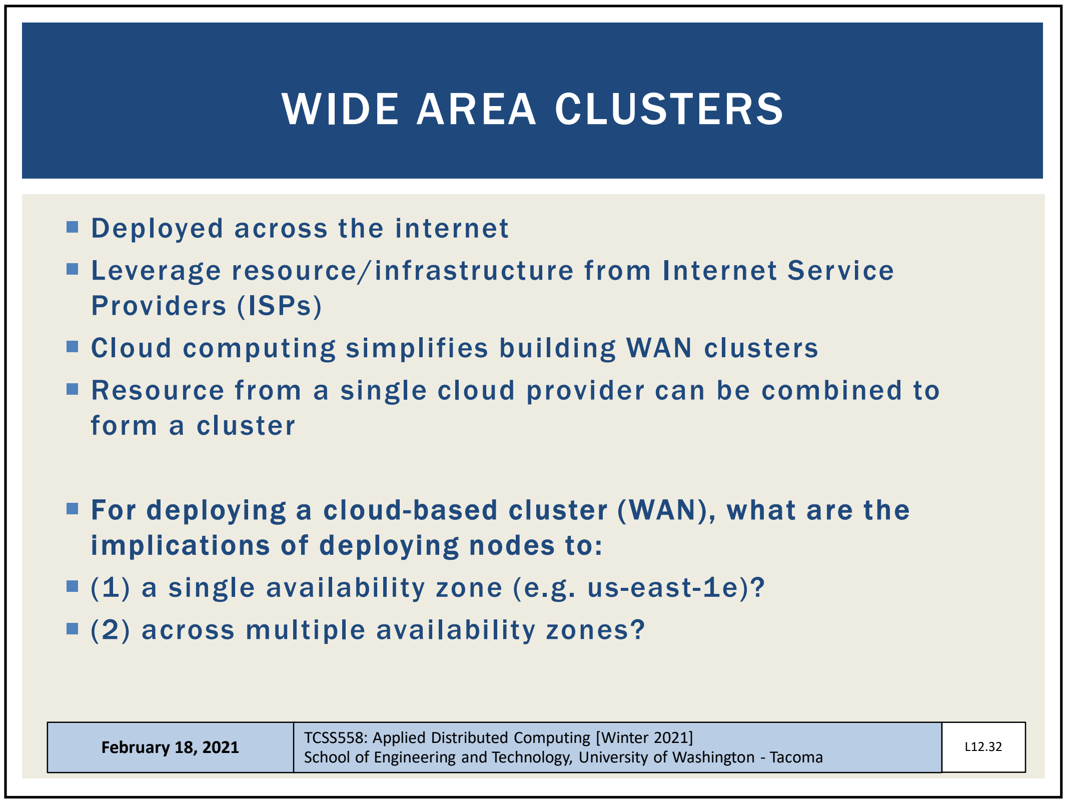
- Front end of three tier architecture (logical switch) provides distribution transparency – hides multiple servers
- Transport-layer switches: switch accepts TCP connection requests, hands off to a server
  - Example: hardware load balancer (F5 networks – Seattle)
  - HW Load balancer - OSI layers 4-7
- Network-address-translation (NAT) approach:
  - All requests pass through switch
  - Switch sits in the middle of the client/server TCP connection
  - Maps (rewrites) source and destination addresses
- Connection hand-off approach:
  - TCP Handoff: switch hands of connection to a selected server



LAN REQUEST DISPATCHING - 2

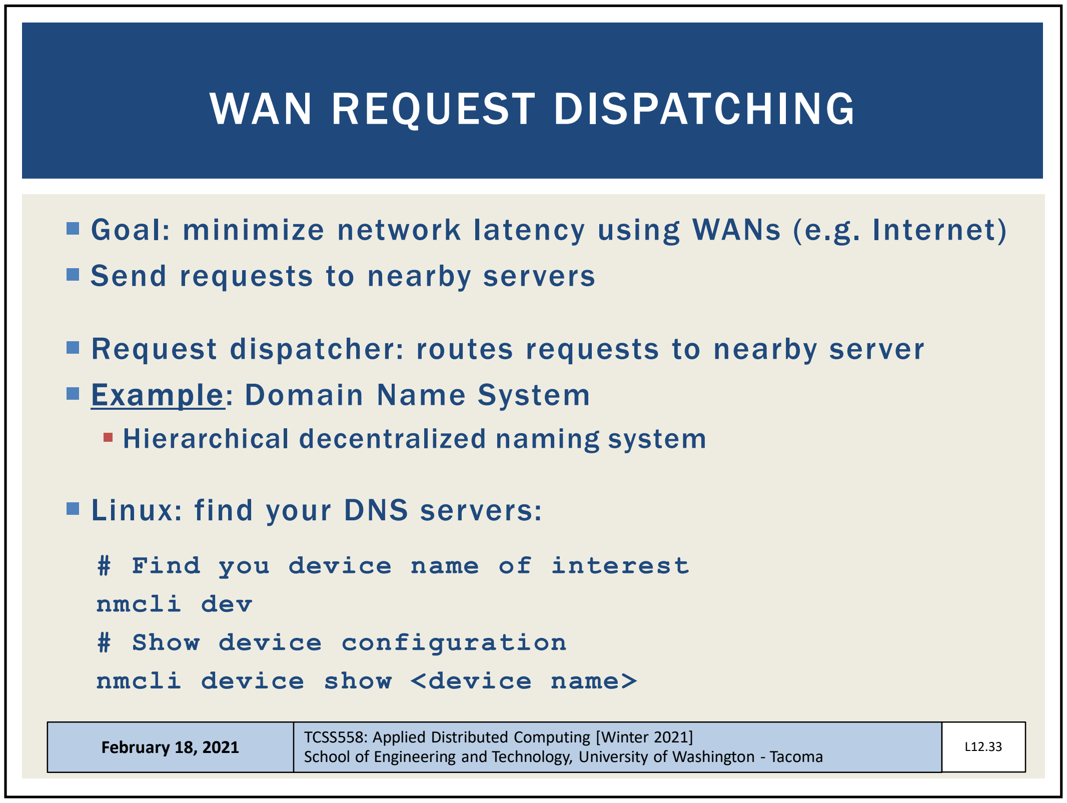
- Who is the best server to handle the request?
- Switch plays important role in distributing requests
- Implements load balancing
- **Round-robin** – routes client requests to servers in a looping fashion
- **Transport-level** – route client requests based on TCP port number
- **Content-aware request distribution** – route requests based on inspecting data payload and determining which server node should process the request





WIDE AREA CLUSTERS

- Deployed across the internet
- Leverage resource/infrastructure from Internet Service Providers (ISPs)
- Cloud computing simplifies building WAN clusters
- Resource from a single cloud provider can be combined to form a cluster
- **For deploying a cloud-based cluster (WAN), what are the implications of deploying nodes to:**
- (1) a single availability zone (e.g. us-east-1e)?
- (2) across multiple availability zones?

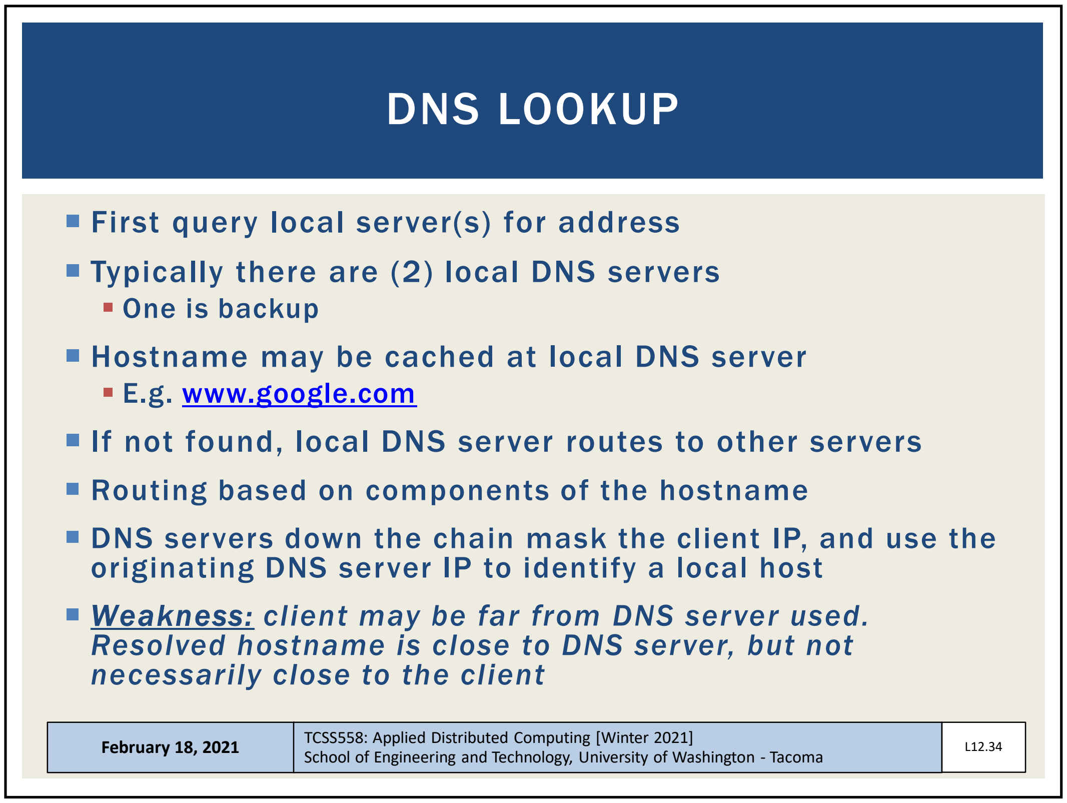


WAN REQUEST DISPATCHING

- Goal: minimize network latency using WANs (e.g. Internet)
- Send requests to nearby servers
- Request dispatcher: routes requests to nearby server
- **Example**: Domain Name System
  - Hierarchical decentralized naming system
- Linux: find your DNS servers:

```
# Find you device name of interest
nmcli dev
# Show device configuration
nmcli device show <device name>
```

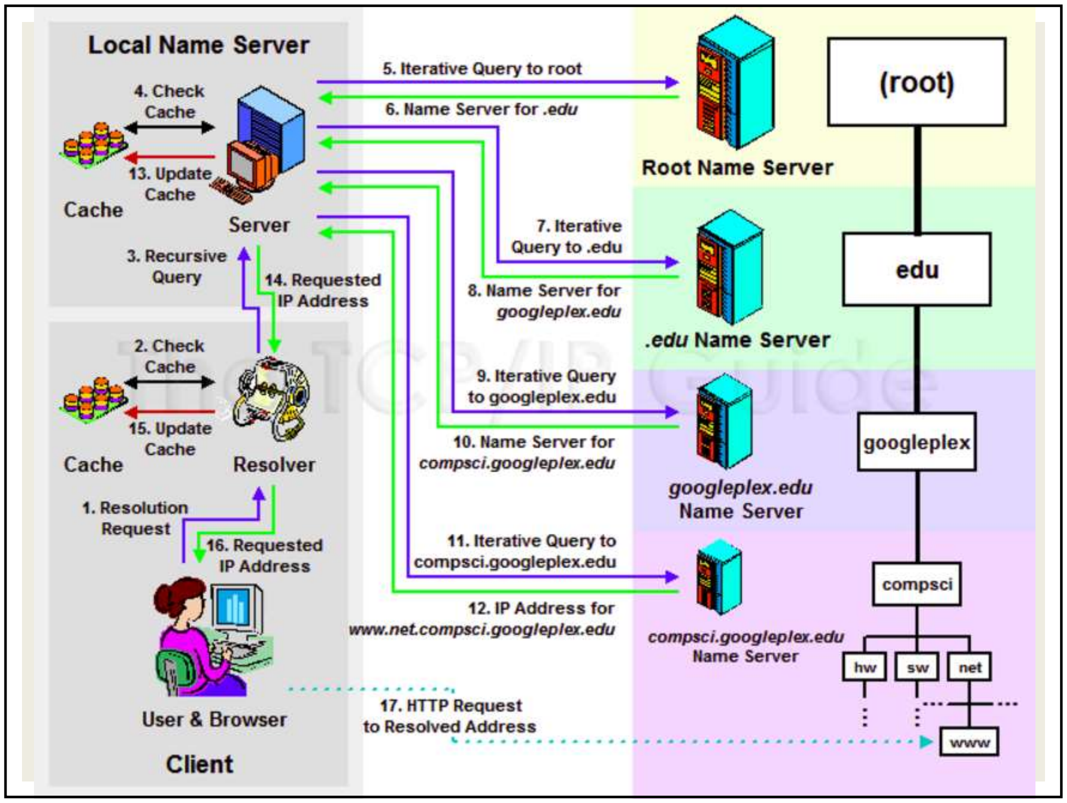
February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.33



DNS LOOKUP

- First query local server(s) for address
- Typically there are (2) local DNS servers
  - One is backup
- Hostname may be cached at local DNS server
  - E.g. www.google.com
- If not found, local DNS server routes to other servers
- Routing based on components of the hostname
- DNS servers down the chain mask the client IP, and use the originating DNS server IP to identify a local host
- ***Weakness:*** *client may be far from DNS server used. Resolved hostname is close to DNS server, but not necessarily close to the client*

February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.34


Local Name Server
4. Check Cache
13. Update Cache
Cache
Server
3. Recursive Query
14. Requested IP Address
2. Check Cache
15. Update Cache
Cache
Resolver
1. Resolution Request
16. Requested IP Address
User & Browser
Client
5. Iterative Query to root
6. Name Server for .edu
7. Iterative Query to .edu
8. Name Server for googleplex.edu
9. Iterative Query to googleplex.edu
10. Name Server for compsci.googleplex.edu
11. Iterative Query to compsci.googleplex.edu
12. IP Address for www.net.compsci.googleplex.edu
17. HTTP Request to Resolved Address
Root Name Server
.edu Name Server
googleplex.edu Name Server
compsci.googleplex.edu Name Server
(root)
edu
googleplex
compsci
hw
sw
net
www
The TCP/IP Guide

DNS: LINUX COMMANDS

- `nslookup <ip addr / hostname>`
- Name server lookup – translates hostname or IP to the inverse

- `traceroute <ip addr / hostname>`
- Traces network path to destination
- By default, output is limited to 30 hops, can be increased

DNS EXAMPLE – WAN DISPATCHING

- Ping www.google.com in WA from wireless network:
  - nslookup: 6 alternate addresses returned, choose (74.125.28.147)
  - Ping 74.125.28.147: Average RTT = **22.458 ms (11 attempts, 22 hops)**
- Ping www.google.com in VA (us-east-1) from EC2 instance:
  - nslookup: 1 address returned, choose 172.217.9.196
  - Ping 172.217.9.196: Average RTT = 1.278 ms (11 attempts, 13 hops)

- From VA EC2 instance, ping WA *www.google* server
- Ping 74.125.28.147: Average RTT 62.349ms (11 attempts, 27 hops)
- Pinging the WA-local server is ~60x slower from VA

- From local wireless network, ping VA us-east-1 google :
- Ping 172.217.9.196: Average RTT=81.637ms (11 attempts, 15 hops)

DNS EXAMPLE – WAN DISPATCHING

- Ping www.google.com in WA from wireless network:
  - nslookup: 6 alternate addresses returned, choose (74.125.28.147)

Latency to ping VA server in WA: ~3.63x

WA client: local-google 22.458ms to VA-google 81.637ms

Latency to ping WA server in VA: ~48.7x

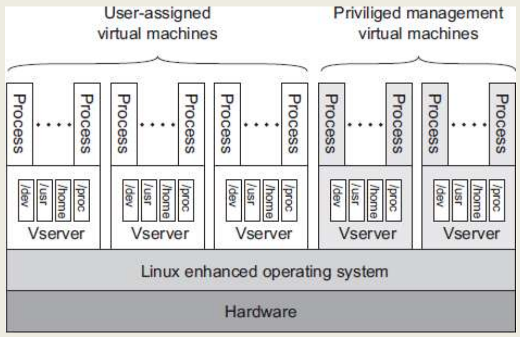
VA client: local-google 1.278ms to WA-google 62.349!

- From local wireless network, ping VA us-east-1 google :
- Ping 172.217.9.196: Average RTT=81.637ms (11 attempts, 15 hops)

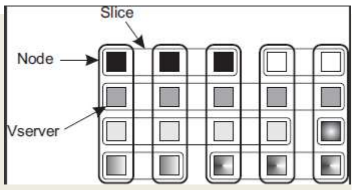
CH 3.2 - EXAMPLE: PLANETLAB

- Unstructured heterogeneous cluster of servers
- Similar to grid but organized as cluster (no grid middleware)
- Testbed established in 2002 for computer networking and distributed systems research
- Organizations share nodes in the cluster

Leverages Linux Vservers
Early "containers"
similar to Docker


User-assigned virtual machines
Priviliged management virtual machines
Process
Process
/dev
/usr
/home
/proc
Vserver
Linux enhanced operating system
Hardware

PLANETLAB - 2



- <u>Slices</u>: set of Vservers running across PlanetLab
- Acts as a virtual server cluster (similar to Amazon VPC)
- <u>Node manager</u>: manages Vservers running on a host
- <u>Slice creation service (SCS)</u>: To create virtual server clusters
- Clients must be <u>slice authorities</u> to create cluster
- <u>Rspec</u>: resource specification
  - Specifies resource requirements for a slice
- <u>Rcap</u>: resource capability
  - Specifies resource capabilities of nodes

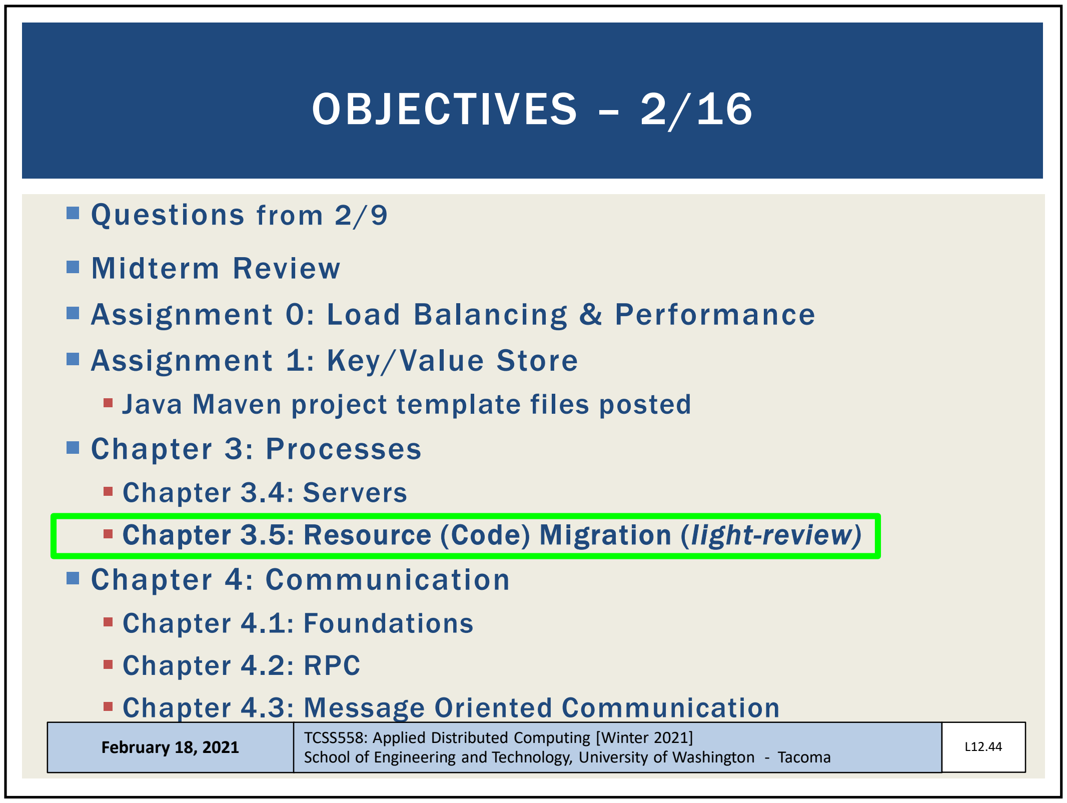
VSERVERS

- Early container based approach
- Vservers share a single operating system kernel
- Primary task is to support a group of processes
- Provides separation of name spaces
- Linux kernel maps process IDs: host OS → Vservers
- Each Vserver has its own set of libraries and file system
- Similar name separation as the "`chroot`" command
- Additional isolation provided to prevent unauthorized access among Vservers directory trees

VSERVERS - 2

- Advantages of Vservers (containers) vs. VMs:
- Simpler resource allocation
- Possible to overbook resources by leveraging dynamic resource allocation - Example: CPU or RAM *(assignment 0, config 2)*
- VMs reserve a block of memory
- Containers can oversubscribe memory
  - Memory not formally reserved
  - Linux kernel shares memory among processes
  - Swap filesystem can use disk as extended RAM
- Memory sharing important for PlanetLab
  - Early nodes had limited memory (e.g. 4 GB)
- Vserver hogging most memory reset when out of swap space

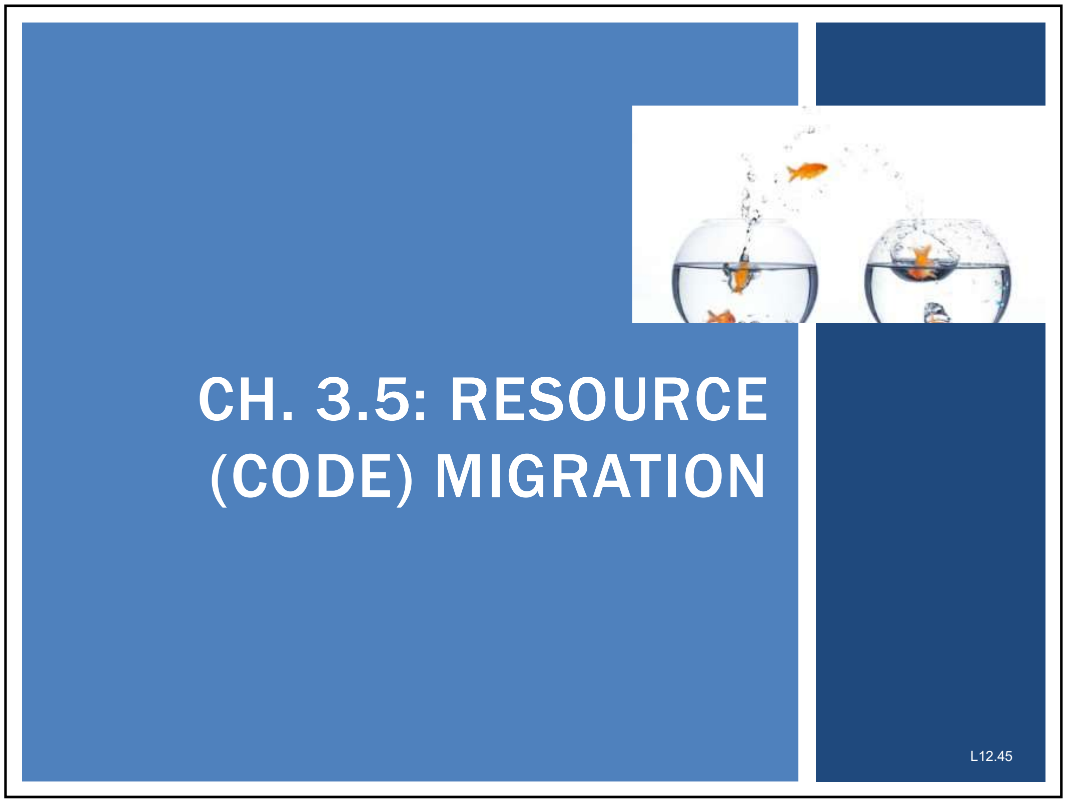
WE WILL RETURN AT
3:02 PM



OBJECTIVES – 2/16

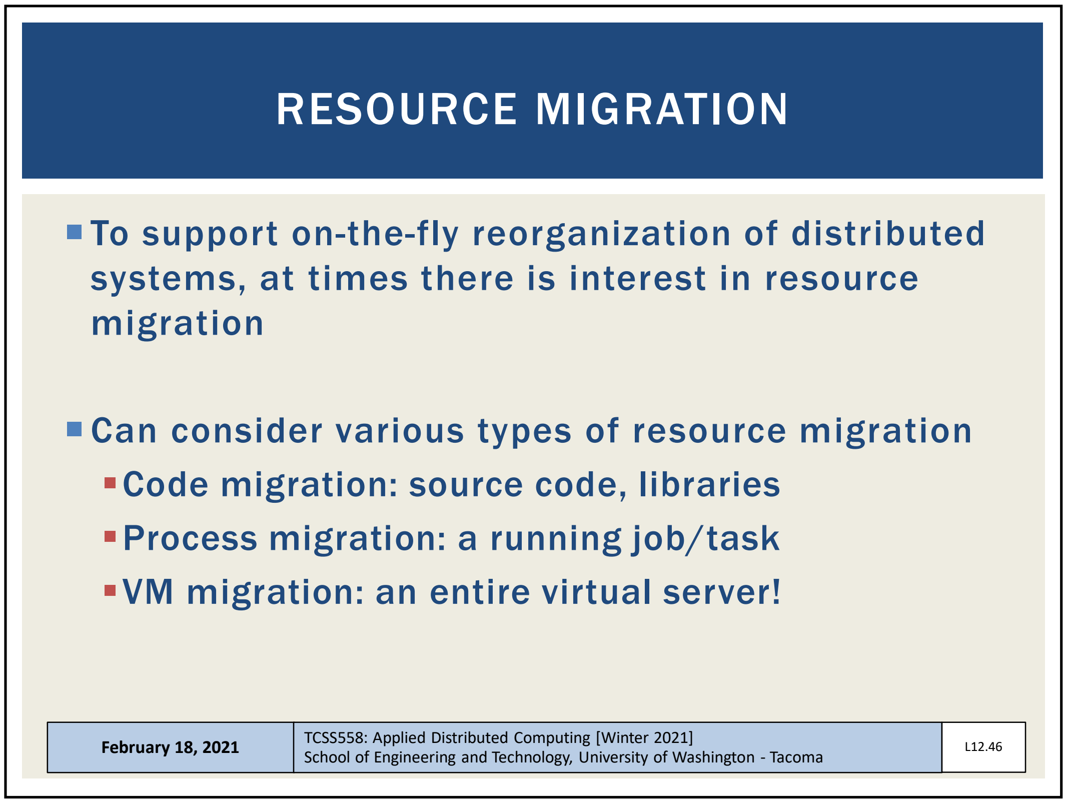
- Questions from 2/9
- Midterm Review
- Assignment 0: Load Balancing & Performance
- Assignment 1: Key/Value Store
  - Java Maven project template files posted
- Chapter 3: Processes
  - Chapter 3.4: Servers
  - **Chapter 3.5: Resource (Code) Migration (*light-review*)**
- Chapter 4: Communication
  - Chapter 4.1: Foundations
  - Chapter 4.2: RPC
  - Chapter 4.3: Message Oriented Communication

February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.44



CH. 3.5: RESOURCE (CODE) MIGRATION

L12.45

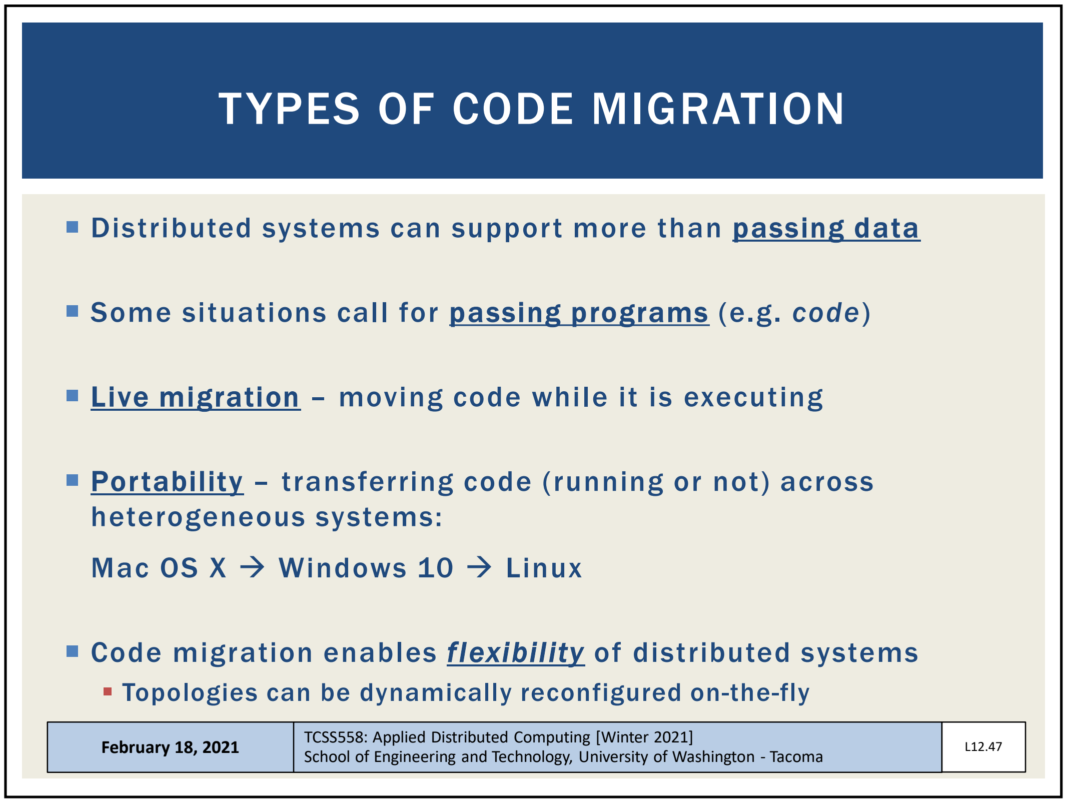


RESOURCE MIGRATION

- To support on-the-fly reorganization of distributed systems, at times there is interest in resource migration

- Can consider various types of resource migration
  - Code migration: source code, libraries
  - Process migration: a running job/task
  - VM migration: an entire virtual server!

February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.46

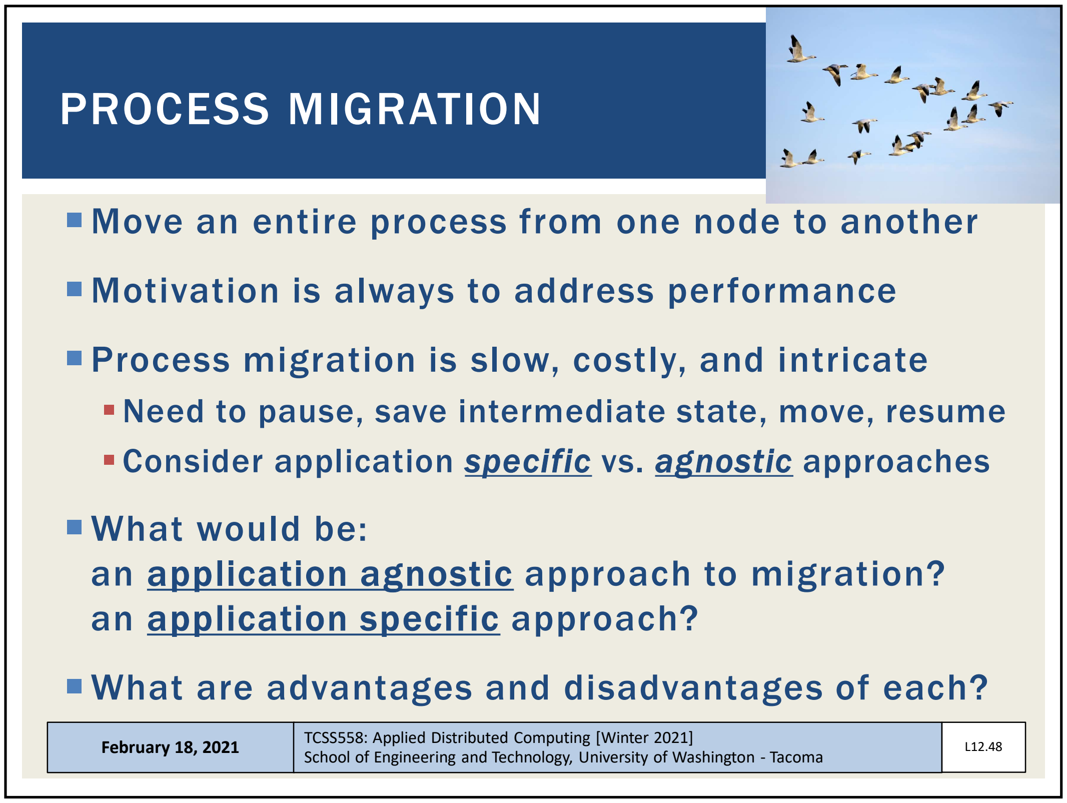


TYPES OF CODE MIGRATION

- Distributed systems can support more than <u>passing data</u>
- Some situations call for <u>passing programs</u> (e.g. *code*)
- <u>Live migration</u> – moving code while it is executing
- <u>Portability</u> – transferring code (running or not) across heterogeneous systems:

  Mac OS X → Windows 10 → Linux
- Code migration enables ***flexibility*** of distributed systems
  - Topologies can be dynamically reconfigured on-the-fly

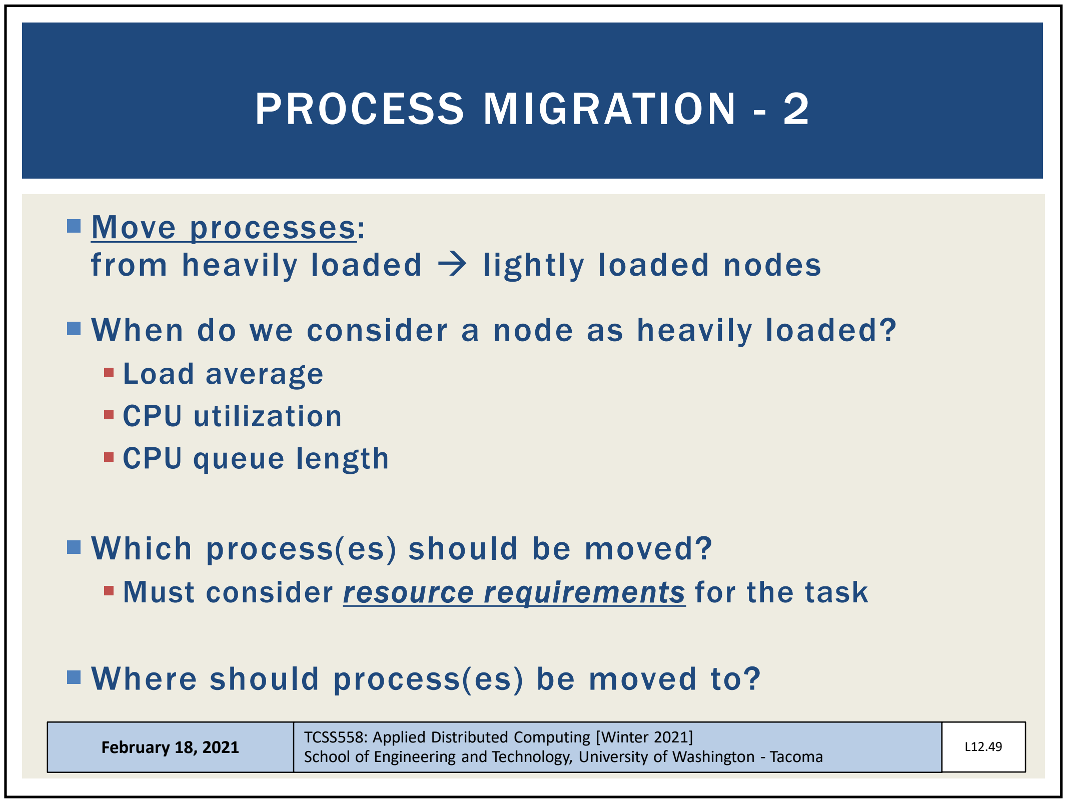
February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.47



PROCESS MIGRATION

- Move an entire process from one node to another
- Motivation is always to address performance
- Process migration is slow, costly, and intricate
  - Need to pause, save intermediate state, move, resume
  - Consider application ***specific*** vs. ***agnostic*** approaches
- What would be:
  an <u>application agnostic</u> approach to migration?
  an <u>application specific</u> approach?
- What are advantages and disadvantages of each?

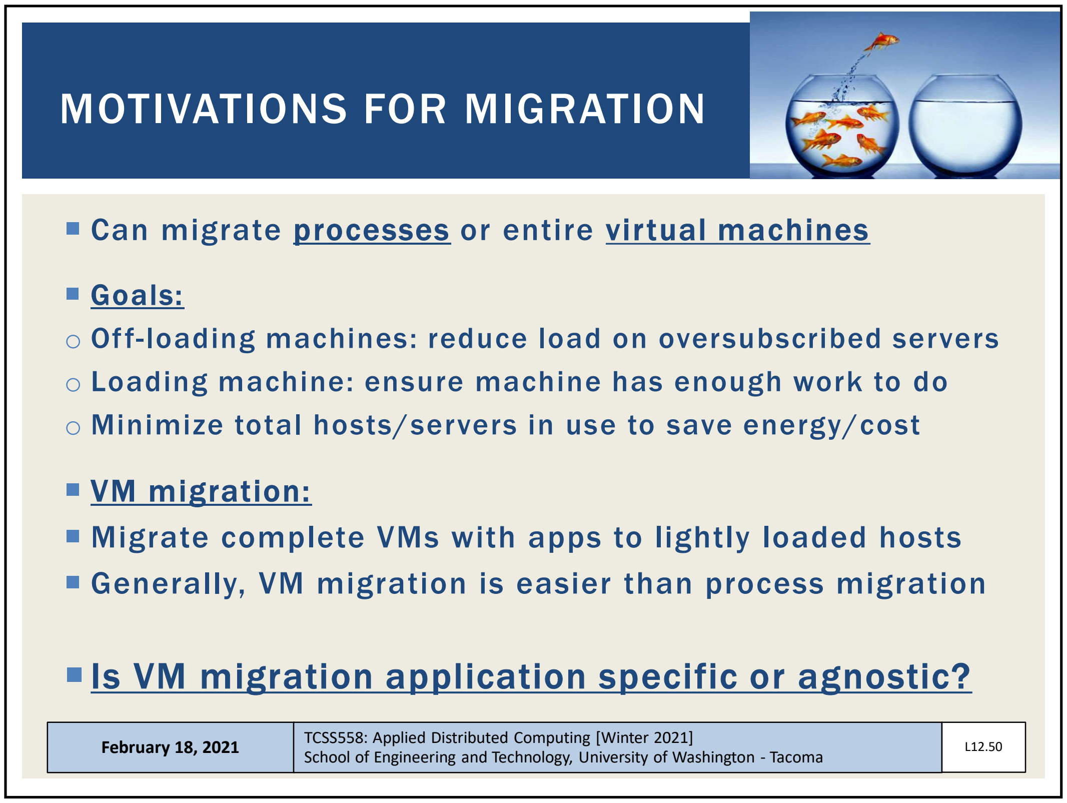
February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.48



PROCESS MIGRATION - 2

- Move processes:
  from heavily loaded → lightly loaded nodes
- When do we consider a node as heavily loaded?
  - Load average
  - CPU utilization
  - CPU queue length
- Which process(es) should be moved?
  - Must consider ***resource requirements*** for the task
- Where should process(es) be moved to?

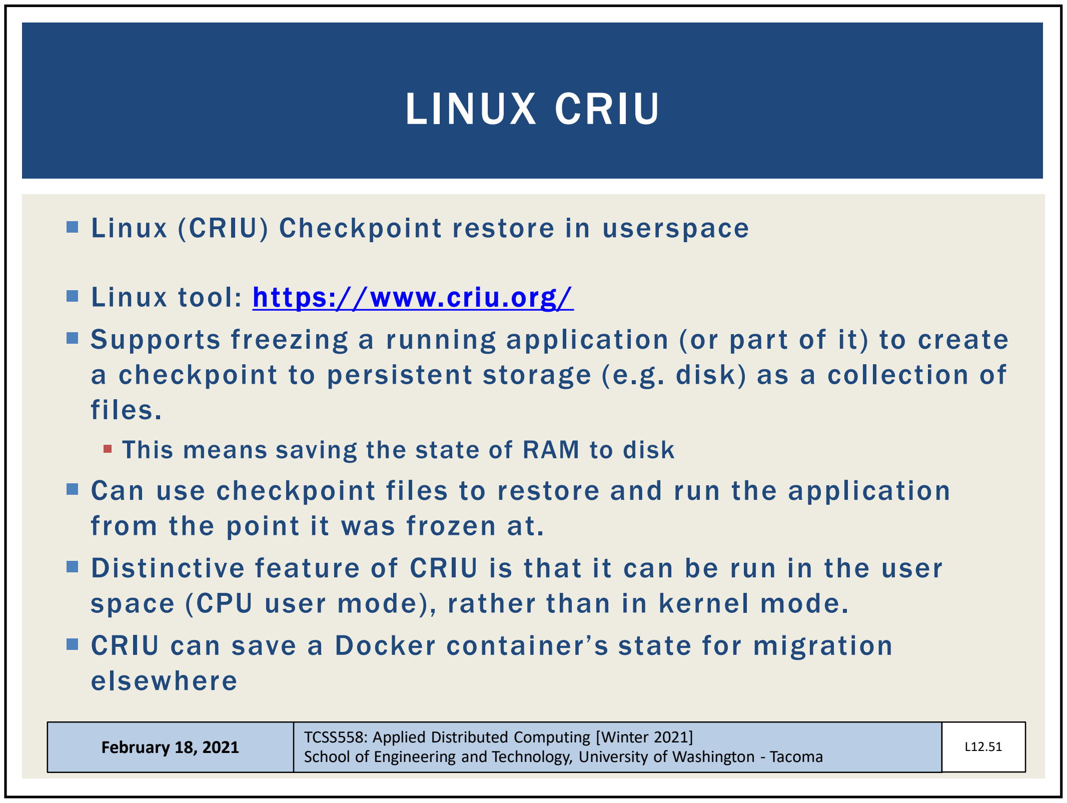
February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.49



MOTIVATIONS FOR MIGRATION

- Can migrate processes or entire virtual machines
- Goals:
  - ○ Off-loading machines: reduce load on oversubscribed servers
  - ○ Loading machine: ensure machine has enough work to do
  - ○ Minimize total hosts/servers in use to save energy/cost
- VM migration:
- Migrate complete VMs with apps to lightly loaded hosts
- Generally, VM migration is easier than process migration
- **Is VM migration application specific or agnostic?**

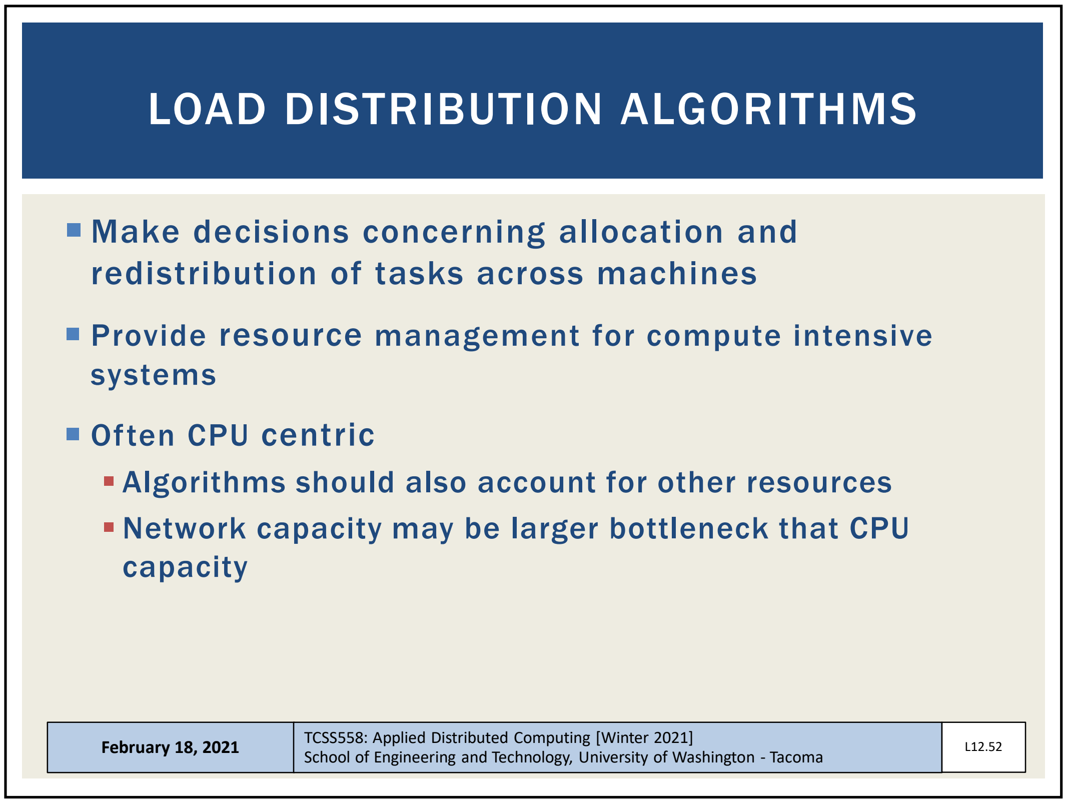
February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.50



LINUX CRIU

- Linux (CRIU) Checkpoint restore in userspace
- Linux tool: https://www.criu.org/
- Supports freezing a running application (or part of it) to create a checkpoint to persistent storage (e.g. disk) as a collection of files.
  - This means saving the state of RAM to disk
- Can use checkpoint files to restore and run the application from the point it was frozen at.
- Distinctive feature of CRIU is that it can be run in the user space (CPU user mode), rather than in kernel mode.
- CRIU can save a Docker container's state for migration elsewhere

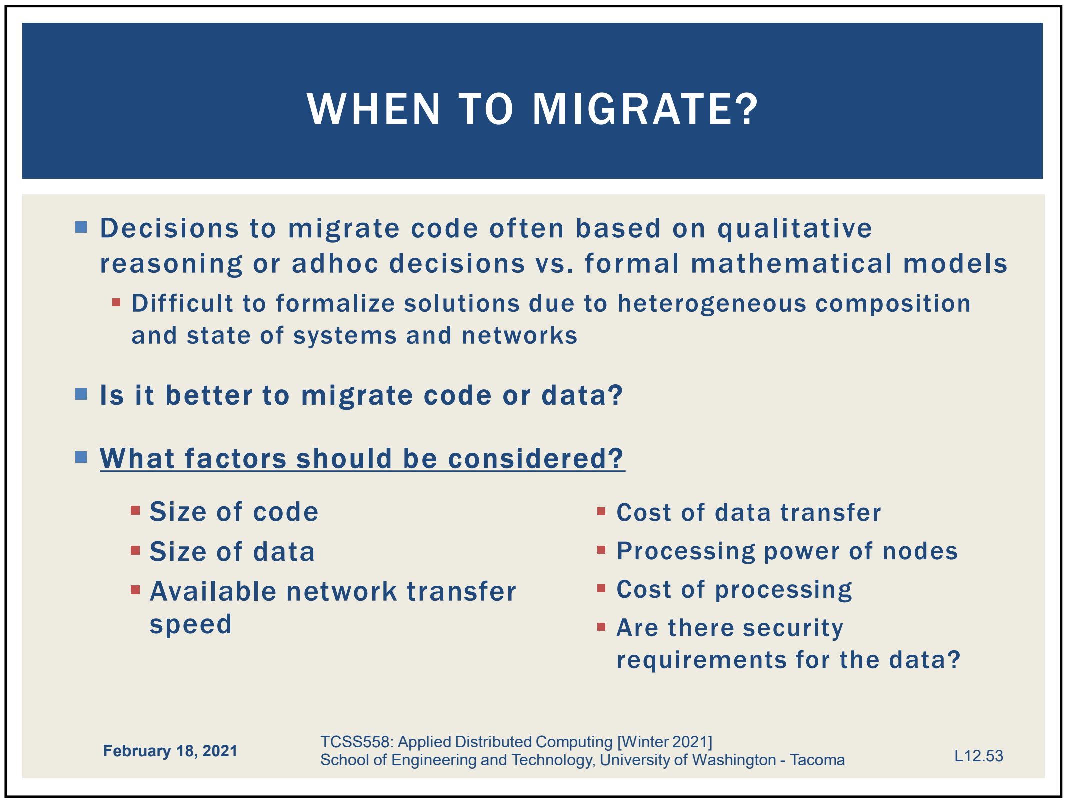
February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.51



LOAD DISTRIBUTION ALGORITHMS

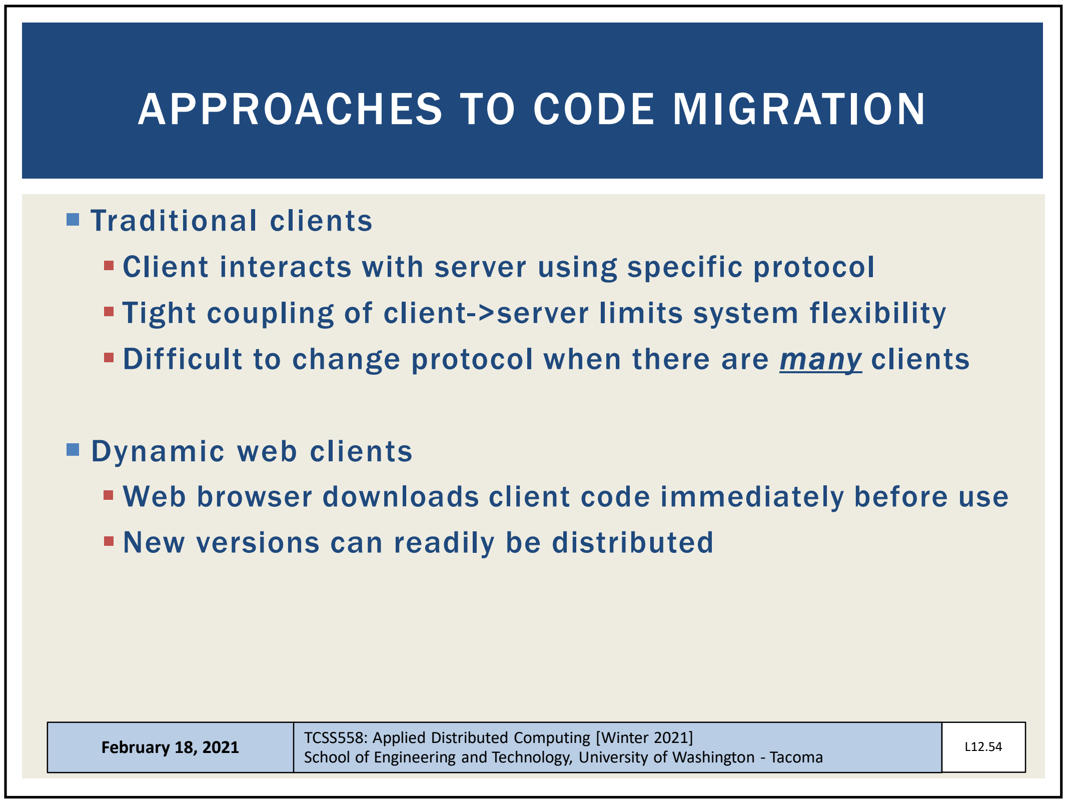
- Make decisions concerning allocation and redistribution of tasks across machines
- Provide resource management for compute intensive systems
- Often CPU centric
  - Algorithms should also account for other resources
  - Network capacity may be larger bottleneck that CPU capacity

February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.52



WHEN TO MIGRATE?

- Decisions to migrate code often based on qualitative reasoning or adhoc decisions vs. formal mathematical models
  - Difficult to formalize solutions due to heterogeneous composition and state of systems and networks
- Is it better to migrate code or data?
- <u>What factors should be considered?</u>
  - Size of code
  - Size of data
  - Available network transfer speed
  - Cost of data transfer
  - Processing power of nodes
  - Cost of processing
  - Are there security requirements for the data?

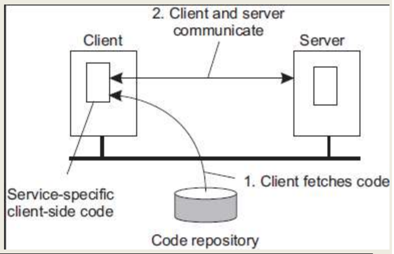


APPROACHES TO CODE MIGRATION

- Traditional clients
  - Client interacts with server using specific protocol
  - Tight coupling of client->server limits system flexibility
  - Difficult to change protocol when there are ***<u>many</u>*** clients
- Dynamic web clients
  - Web browser downloads client code immediately before use
  - New versions can readily be distributed

DYNAMIC WEB CLIENTS

- Advantages
  - Client code loaded in as necessary
  - Discarded when no longer needed
  - Can easily change the client/server protocol
- Disadvantages
  - Security: we have to trust the code
  - Downloading client requires network bandwidth & time


2. Client and server communicate
Client
Server
1. Client fetches code
Service-specific client-side code
Code repository

CODE MIGRATION

- Sender-initiated: (upload the code)… e.g. Github
- Receiver-initiated: (download the code)… e.g. web browser
- Remote cloning
  - Produce a copy of the process on another machine while parent runs

CODE MIGRATION - 2

- What is migrated?
  - ***Code*** segment
  - ***Resource*** segment (device info)
  - ***Execution*** segment (process info: data, state, stack, PC)
- **Weak mobility**
  - Only ***code*** segment, no state
  - Code always restarts
- **Strong mobility**
  - ***Code*** + ***execution*** segment
  - Process stopped, state saved, moved, resumed
  - Represents true ***process migration***

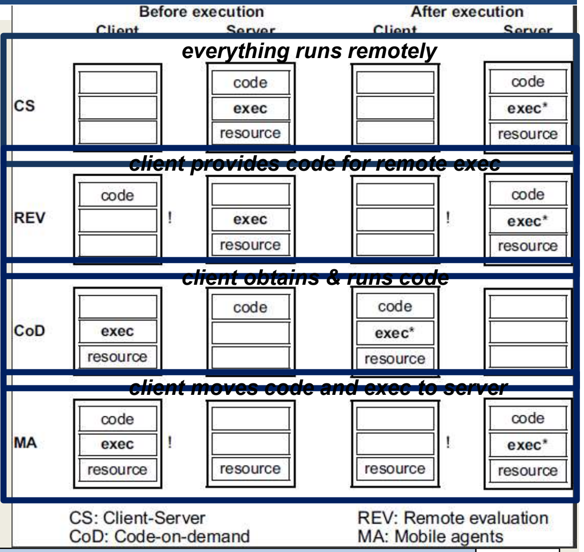
CODE MOBILITY TYPES

* indicates what is modified

- CS: Client-Server
- REV: Remote Evaluation
- CoD: Code-on-demand
- MA: Mobile agents

- Where does state get modified?

- State is stored in **exec**



MIGRATION OF HETEROGENEOUS SYSTEMS

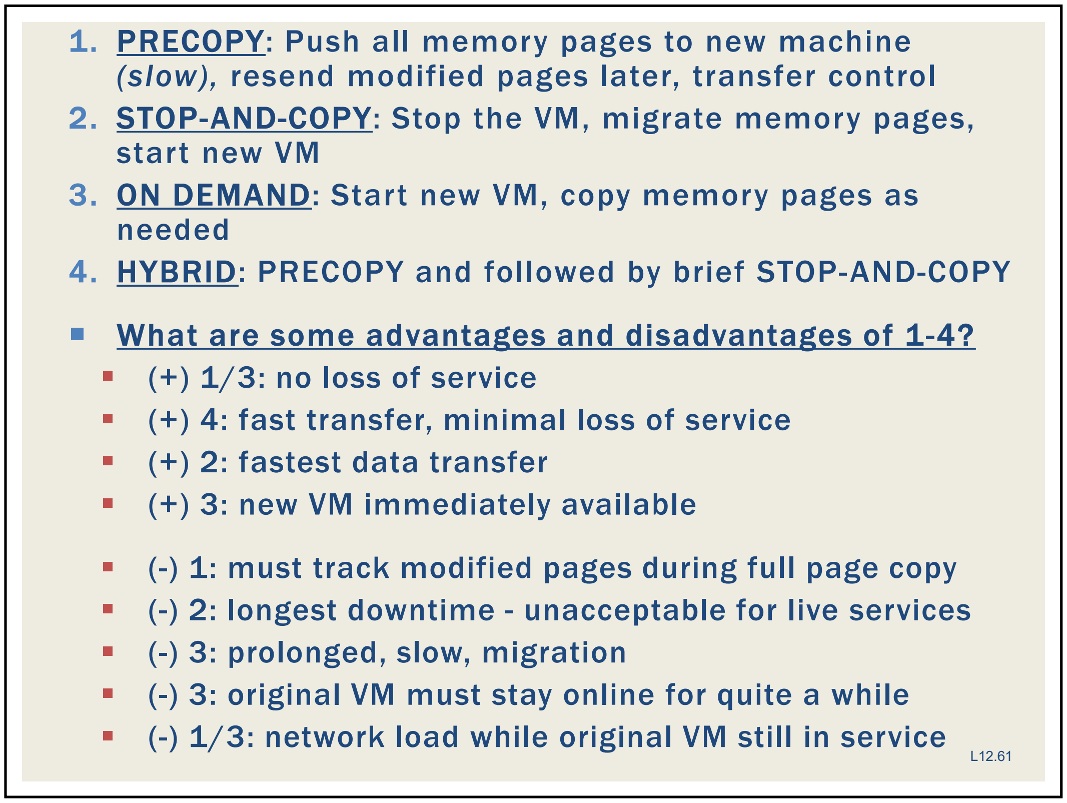
- Assumption: code will always work at new node
- Invalid if node architecture is different (*heterogeneous*)
- What approaches are available to migrate code across heterogeneous systems?
- Intermediate code
  - 1970s Pascal: generate machine-independent intermediate code
  - Programs could then run anywhere
  - **Today:** web languages: Javascript, Java
- VM Migration

VIRTUAL MACHINE MIGRATION

- Four approaches:

1. **PRECOPY**: Push all memory pages to new machine *(slow)*, resend modified pages later, transfer control
2. **STOP-AND-COPY**: Stop the VM, migrate memory pages, start new VM
3. **ON DEMAND**: Start new VM, copy memory as needed
4. **HYBRID**: PRECOPY followed by brief STOP-AND-COPY

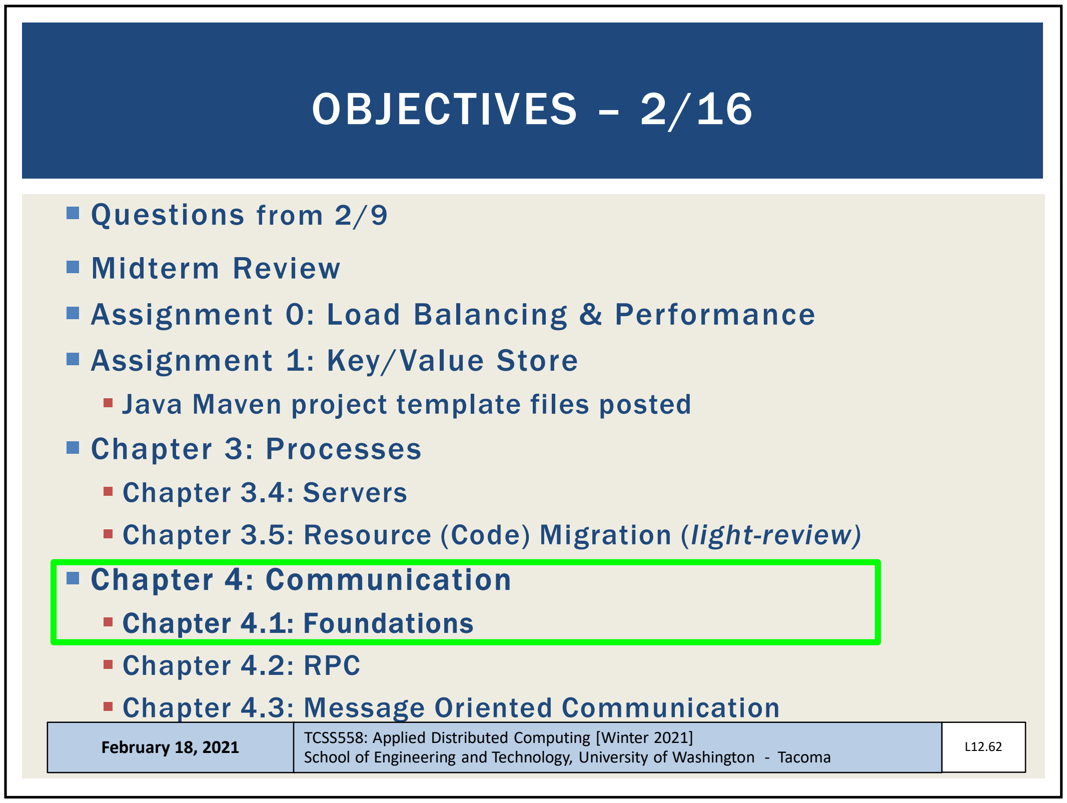
- **What are some advantages and disadvantages of 1-4?**



1. <u>**PRECOPY**</u>: Push all memory pages to new machine *(slow)*, resend modified pages later, transfer control
2. <u>**STOP-AND-COPY**</u>: Stop the VM, migrate memory pages, start new VM
3. <u>**ON DEMAND**</u>: Start new VM, copy memory pages as needed
4. <u>**HYBRID**</u>: PRECOPY and followed by brief STOP-AND-COPY

- <u>What are some advantages and disadvantages of 1-4?</u>
  - (+) 1/3: no loss of service
  - (+) 4: fast transfer, minimal loss of service
  - (+) 2: fastest data transfer
  - (+) 3: new VM immediately available
  - (-) 1: must track modified pages during full page copy
  - (-) 2: longest downtime - unacceptable for live services
  - (-) 3: prolonged, slow, migration
  - (-) 3: original VM must stay online for quite a while
  - (-) 1/3: network load while original VM still in service

L12.61



OBJECTIVES – 2/16

- Questions from 2/9
- Midterm Review
- Assignment 0: Load Balancing & Performance
- Assignment 1: Key/Value Store
  - Java Maven project template files posted
- Chapter 3: Processes
  - Chapter 3.4: Servers
  - Chapter 3.5: Resource (Code) Migration (*light-review*)
- Chapter 4: Communication
  - Chapter 4.1: Foundations
  - Chapter 4.2: RPC
  - Chapter 4.3: Message Oriented Communication

February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.62

CH. 4 COMMUNICATION

CHAPTER 4

- 4.1 Foundations
  - Protocols
  - Types of communication
- 4.2 Remote procedure call
- 4.3 Message-oriented communication
  - Socket communication
  - Messaging libraries
  - Message-Passing Interface (MPI)
  - Message-queueing systems
  - Examples
- 4.4 Multicast communication
  - Flooding-based multicasting
  - Gossip-based data dissemination

Reviews and builds on content from Ch. 2/3

CH. 4.1: FOUNDATIONS

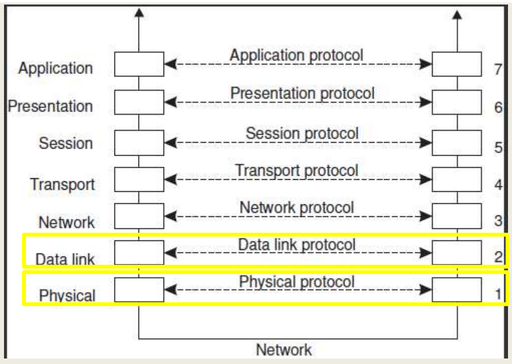
LAYERED PROTOCOLS

- Distributed systems lack shared memory
- All distributed system communication is based on sending and receiving low-level messages
  - P → Q
- <u>O</u>pen <u>S</u>ystems <u>I</u>nterconnection Reference Model (OSI Model)
  - Open systems communicate with any other open system
  - Standards govern format, contents, meaning of messages
  - Formalization of rules forms a **<u>communication protocol</u>**

LAYERED PROTOCOLS - 2

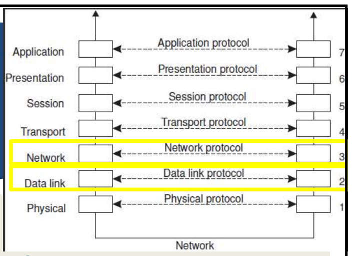
- Protocols provide a communication service
- Two service types:
  - **Connection-oriented**: sender/receiver establish connection, negotiate parameters of the protocol, close connection when done
  - Physical example: telephone
  - **Connectionless**: No setup. Sender sends. Receiver receives.
  - Physical example: Mailing a letter

OSI MODEL REVISITED



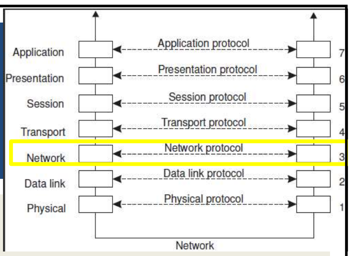
- Physical layer: just sends bits → … 0 0 0 1 0 1 1 0 1 1 …
- Data link layer: Groups bits into frames
  - Provides error correction via *checksum*
  - Special bit pattern at start/end of frame

OSI MODEL - 2



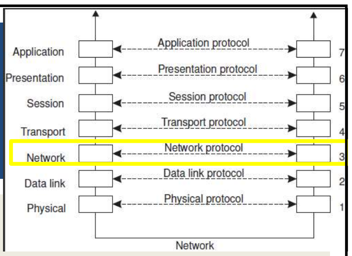
- Data link layer:
  - <u>Checksum</u>: computed by adding all bytes in frame in particular way
  - Added to message
  - Receiver removes checksum, recomputes checksum, and compares
  - If receiver and sender agree, frame is considered correct
  - Receiver can request failed frames to be resent
  - Frames assigned sequence numbers *<u>in the header</u>*
- Network layer:
  - Sometimes referred to as the *Internet layer*
  - On WANs sending msgs between client/server requires routing
  - Provides addressing using IPV4 (32-bit), IPV6 (64-bit)

OSI MODEL - 3



- Network layer:
  - Helps with routing network traffic
  - Shortest route (# of hops) may not be the best route
  - Minimizing delay (latency) is paramount
  - Routing algorithms: use long-term average network conditions, or try to adapt to changing conditions
  - ICMP Protocol: Internet Control Message Protocol
  - Not typically for sending data, used for diagnostic/control purposes
  - ICMP Examples: (*ping*, *traceroute*)

OSI MODEL - 4



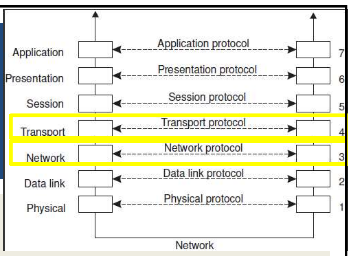
- Internet Control Message Protocol (ICMP)
  - 8 bytes header: 4 fixed, 4 variable

ICMP Header Format

Offsets	Octet	0								1								2								3							
Octet	Bit	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	0	Type								Code								Checksum															
4	32	Rest of Header																															

- Example message types:
- 0- echo reply (PING), 3- destination unreachable, 4- source quench (congestion control), 5- redirect message, 8- echo request (PING), 9- router advertisement
- Others: 10 (router solicitation), 11 (time exceeded), 12 (parameter problem), 13 (timestamp), 15 (info request), 16 (info reply), 17 (address mask request), 18 (address mask reply), 30-39 (traceroute), 40 (security failures), 42 (ext echo request)…255

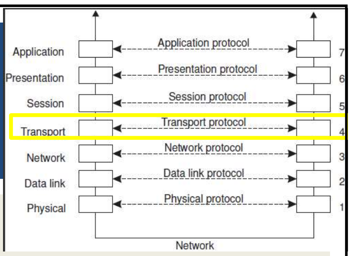
OSI MODEL - 5



- Transport layer:
  - Provides reliable connections
  - Reorganizes packets arriving out of sequence
  - Requests delivery of missing packets

1. Breaks application layer protocol messages into pieces to transmit
2. Assigns messages sequence numbers
3. Sends all messages

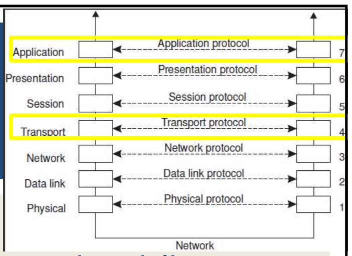
OSI MODEL - 6



- Transport layer provides an infallible "message pipe"
  - Put messages in
  - Always come out undamaged, in correct order

- Transport layer protocols:
  - TCP: Transmission Control Protocol (connection-oriented)
  - UDP: Universal Datagram Protocol (connectionless)

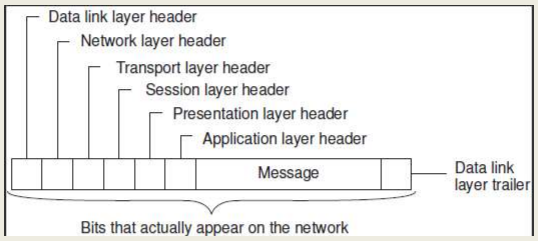
OSI MODEL - 7



- Other transport protocols
  - Real-time transport protocol (RTP): real-time data, no data delivery guarantee
  - Streaming Control Transmission Protocol (SCTP): alternative to TCP

- Higher-level protocols:
- Session layer: mechanisms for opening, closing, managing session between communicating processes
- Presentation layer: deals with syntactical meaning of messages
  - Presentation services convert data among formats, for example:
  - from extended binary coded decimal interchange code (EBCDIC) to ASCII
- Application layer: protocols that don't fit into other layers
  - Many protocols: FTP, SFTP, HTTP, etc. etc.

OSI MODEL - 8

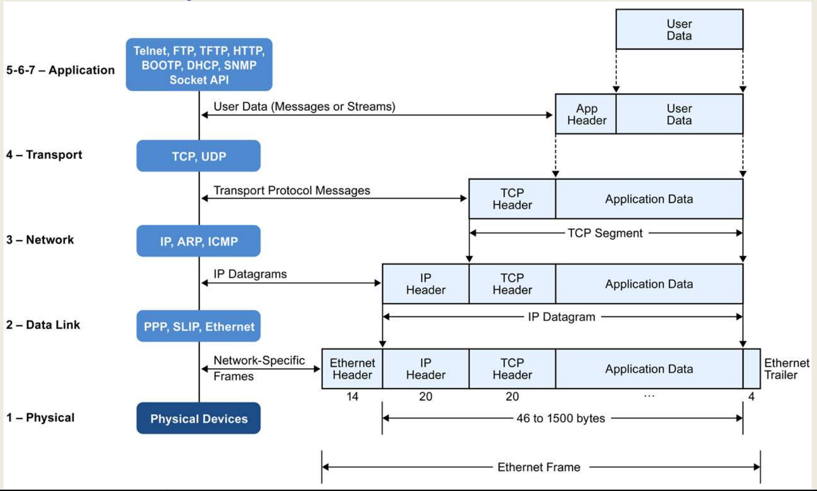


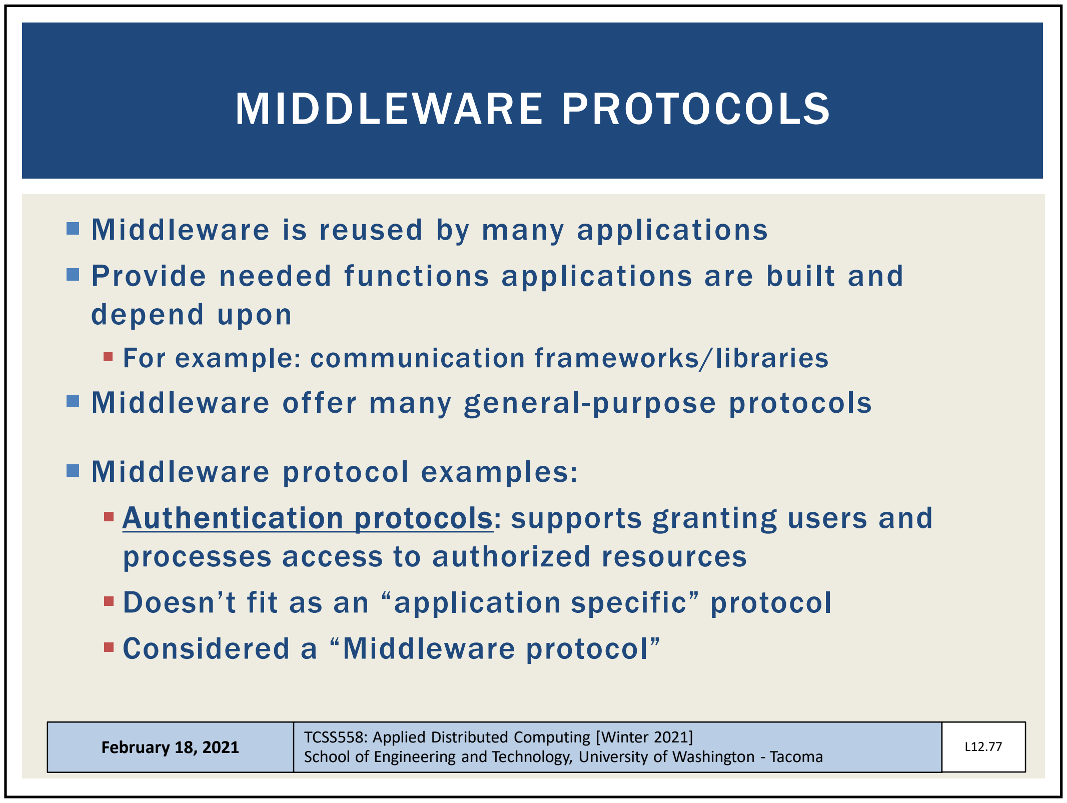
- Each OSI layer contributes overhead bits to the message
- Layers append data to front (and maybe end) of the message
- Receiver strips off headers as the message goes up the OSI model stack:

  physical → data-link → network → transport → application

PROTOCOL STACK

- Collection of layers used for communication from OSI model

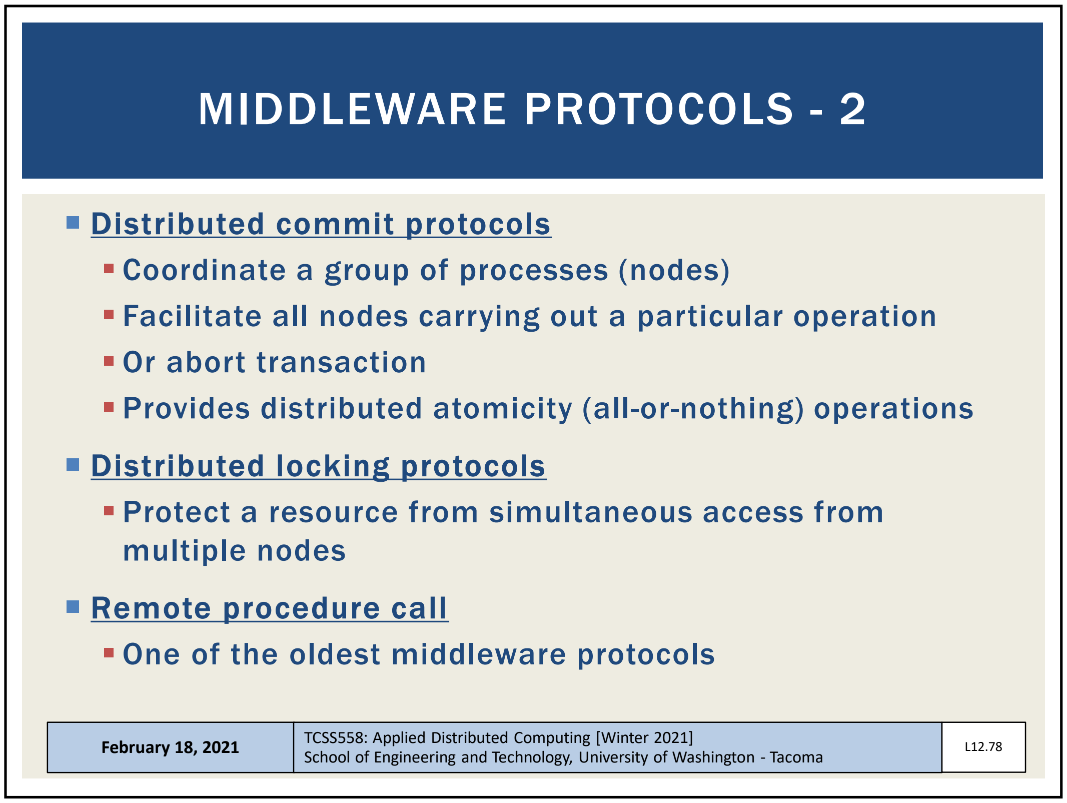




MIDDLEWARE PROTOCOLS

- Middleware is reused by many applications
- Provide needed functions applications are built and depend upon
  - For example: communication frameworks/libraries
- Middleware offer many general-purpose protocols
- Middleware protocol examples:
  - Authentication protocols: supports granting users and processes access to authorized resources
  - Doesn't fit as an "application specific" protocol
  - Considered a "Middleware protocol"

February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.77



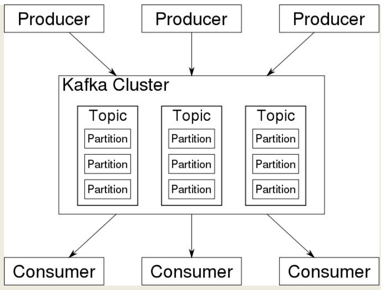
MIDDLEWARE PROTOCOLS - 2

- Distributed commit protocols
  - Coordinate a group of processes (nodes)
  - Facilitate all nodes carrying out a particular operation
  - Or abort transaction
  - Provides distributed atomicity (all-or-nothing) operations
- Distributed locking protocols
  - Protect a resource from simultaneous access from multiple nodes
- Remote procedure call
  - One of the oldest middleware protocols

February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.78

MIDDLEWARE PROTOCOLS - 3

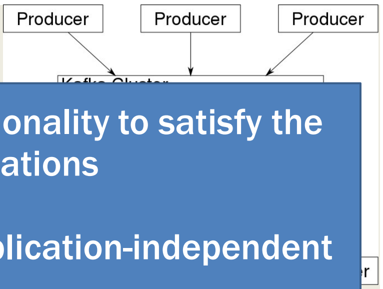
- **Message queueing services**
  - Support synchronization of data streams
  - Transfer real-time data
  - Distributed and scalable implementation



- **Multicast services**
  - Scale communication to thousands of receivers spread across the Internet

MIDDLEWARE PROTOCOLS - 3

- **Message queueing services**
  - Support synchronization of data

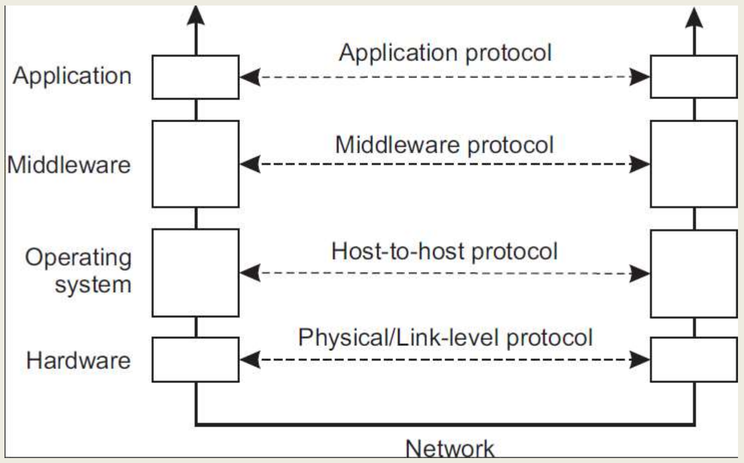


KEY: middleware protocols offer functionality to satisfy the software requirements of *many* applications

Middleware functions are general, application-independent in nature

Functions are so commonly needed they are offered in reusable frameworks / libraries

ADAPTED REFERENCE MODEL



Combines network and transport

Physical and Data link

TYPES OF COMMUNICATION

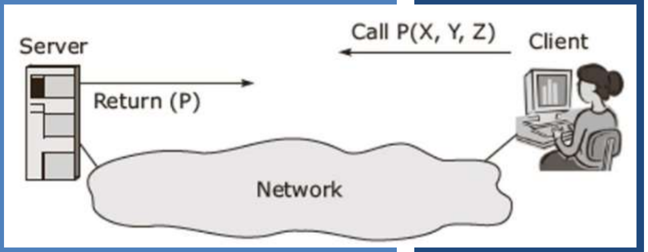
- Persistent communication
  - Message submitted for transmission is stored by communication middleware as long as it takes to deliver it
  - Example: email system (SMTP)
  - Receiver can be offline when message sent
  - Temporal decoupling (delayed message delivery)
- Transient communication
  - Message stored by middleware only as long as sender/receiver applications are running
  - If recipient is not active, message is dropped
  - Transport level protocols typically are transient (*no msg storage*)
- What OSI protocol level is the SMTP Protocol?

TYPES OF COMMUNICATION - 2

- Asynchronous communication
  - Client does not block, continues doing other work
- Synchronous communication
  - Client blocks and waits
- Three types of blocking
  1. Until middleware notifies it will take over delivering ***request***
  2. Sender may block until ***request*** has been delivered
  3. Sender waits until ***request*** is processed and result is returned
- Persistence + synchronization (blocking)
  - Common scheme for message-queueing systems
  - Block until message delivered to queue
- **Consider each type of blocking (1, 2, 3). Are these modes connectionless (UDP)? connection-oriented (TCP)?**

OBJECTIVES – 2/16

- Questions from 2/9
- Midterm Review
- Assignment 0: Load Balancing & Performance
- Assignment 1: Key/Value Store
  - Java Maven project template files posted
- Chapter 3: Processes
  - Chapter 3.4: Servers
  - Chapter 3.5: Resource (Code) Migration (*light-review)*
- Chapter 4: Communication
  - Chapter 4.1: Foundations
  - **Chapter 4.2: RPC**
  - Chapter 4.3: Message Oriented Communication



CH. 4.2: RPC

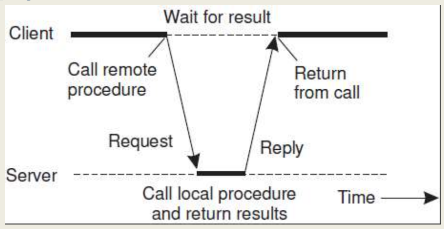
L12.85

RPC – REMOTE PROCEDURE CALL

- In a nutshell,
- Allow programs to call procedures on other machines
- Process on machine A calls procedure on machine B
- Calling process on machine A is suspended
- Execution of the called procedure takes place on machine B
- Data transported from caller (A) to provider (B) and back (A).
- No message passing is visible to the programmer
- Distribution transparency: make remote procedure call look like a local one
- `newlist = append(data, dbList)`

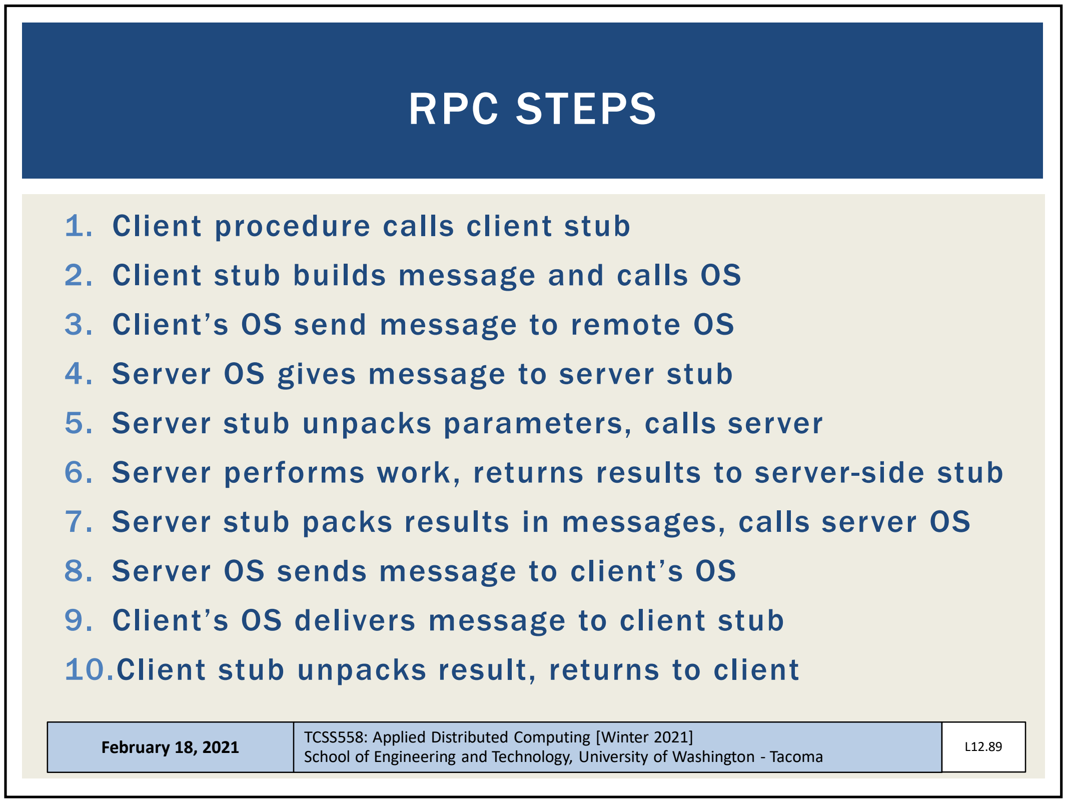
RPC - 2

- Transparency enabled with client and server "stubs"
- Client has "stub" implementation of the server-side function
- Interface exactly same as server side
- But client DOES NOT HAVE THE IMPLEMENTATION
- Client stub: packs parameters into message, sends ***request*** to server. Call blocks and waits for reply
- Server stub: transforms incoming ***request*** into local procedure call
- Blocks to wait for ***reply***
- Server stub unpacks ***request***, calls server procedure
- ***It's as if the routine were called locally***



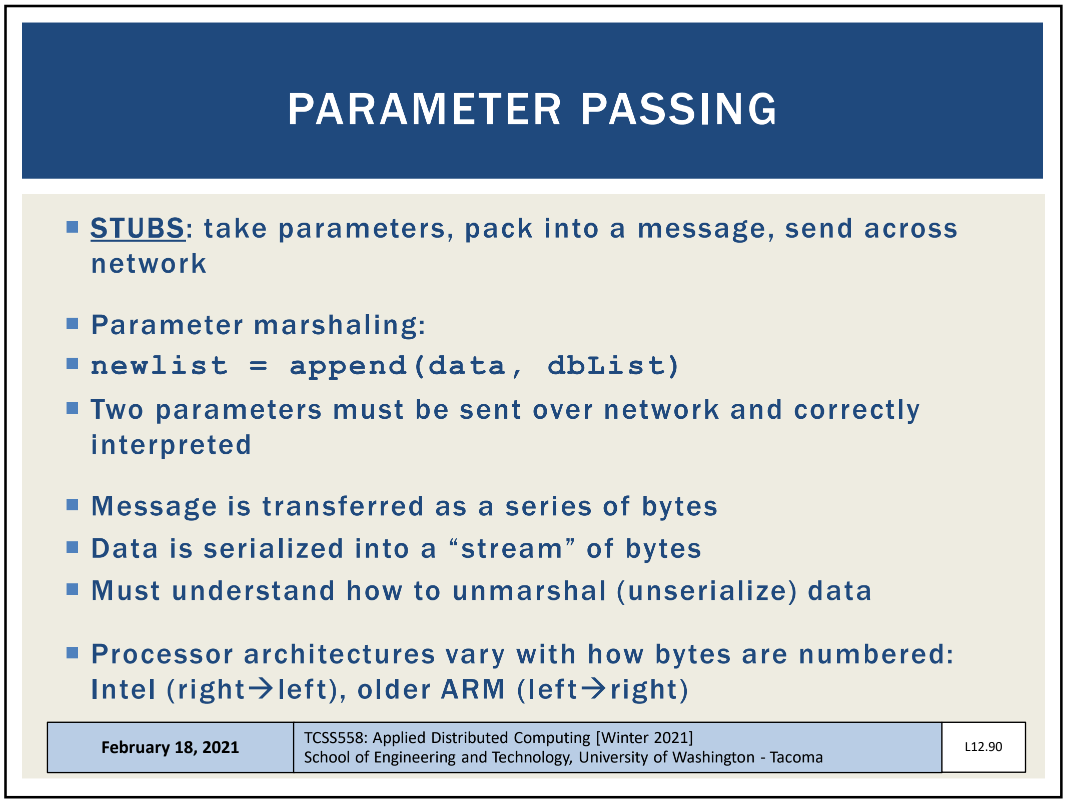
RPC - 3

- Server packs procedure ***results*** and sends back to client.
- Client "***request***" call unblocks and data is unpacked
- Client can't tell method was called remotely over the network… ***except for network latency…***
- Call abstraction enables clients to invoke functions in alternate languages, on different machines
- Differences are handled by the RPC "framework"



RPC STEPS

1. Client procedure calls client stub
2. Client stub builds message and calls OS
3. Client's OS send message to remote OS
4. Server OS gives message to server stub
5. Server stub unpacks parameters, calls server
6. Server performs work, returns results to server-side stub
7. Server stub packs results in messages, calls server OS
8. Server OS sends message to client's OS
9. Client's OS delivers message to client stub
10. Client stub unpacks result, returns to client

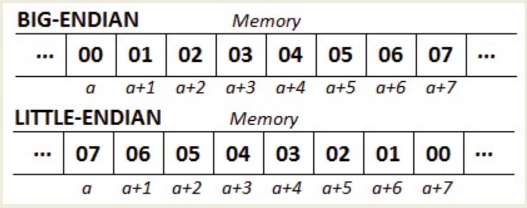


PARAMETER PASSING

- **STUBS**: take parameters, pack into a message, send across network
- Parameter marshaling:
- `newlist = append(data, dbList)`
- Two parameters must be sent over network and correctly interpreted
- Message is transferred as a series of bytes
- Data is serialized into a "stream" of bytes
- Must understand how to unmarshal (unserialize) data
- Processor architectures vary with how bytes are numbered: Intel (right→left), older ARM (left→right)

RPC: BYTE ORDERING

- Big-Endian: write bytes left to right (ARM)
- Little-endian: write bytes right to left (Intel)
- Networks: typically transfer data in Big-Endian form
- Solution: transform data to machine/network independent format
- Marshaling/unmarshaling: transform data to neutral format



BIG-ENDIAN Memory

...	00	01	02	03	04	05	06	07	...
	a	a+1	a+2	a+3	a+4	a+5	a+6	a+7	

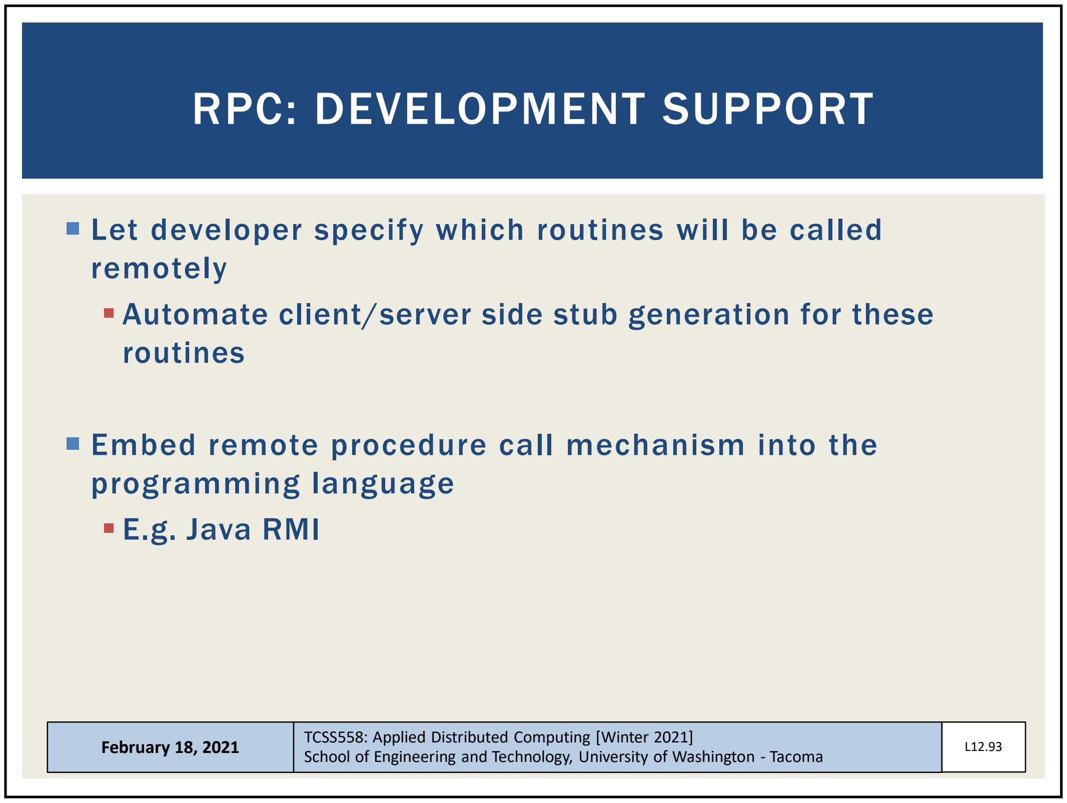
LITTLE-ENDIAN Memory

...	07	06	05	04	03	02	01	00	...
	a	a+1	a+2	a+3	a+4	a+5	a+6	a+7	

RPC: PASS-BY-REFERENCE

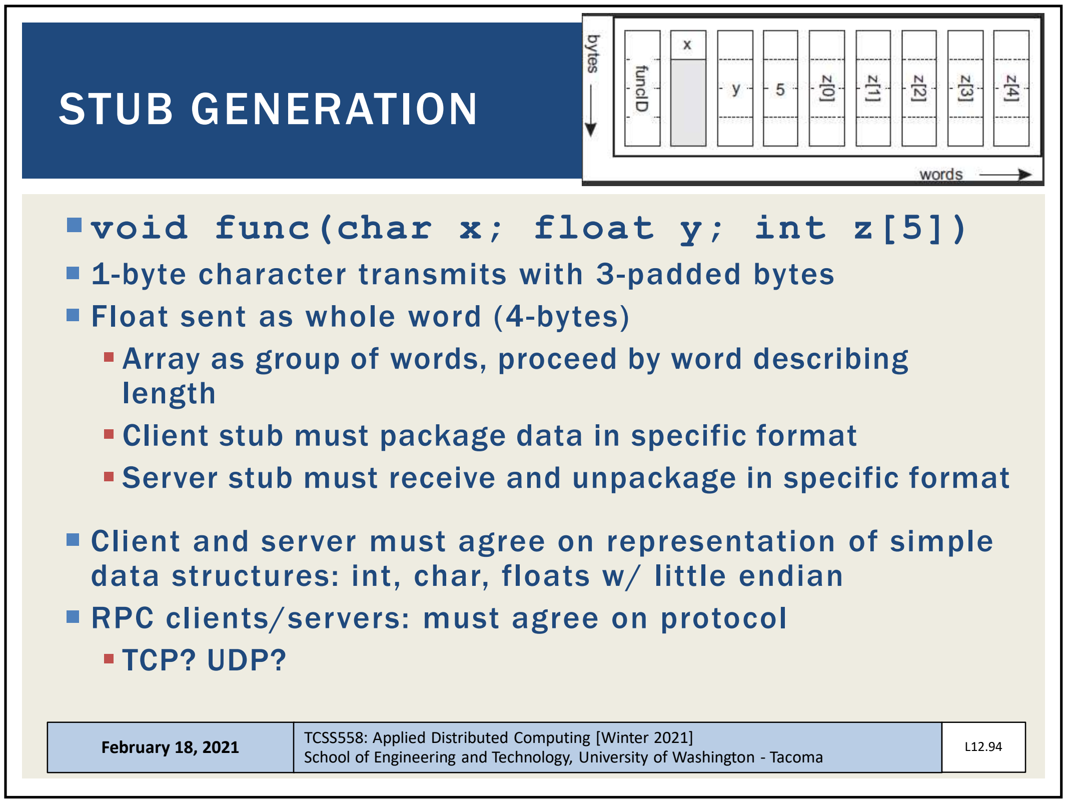
- Passing by value is straightforward
- Passing by reference is challenging
- Pointers only make sense on local machine owning the data
- Memory space of client and server are different

- Solutions to RPC pass-by-reference:
1. Forbid pointers altogether
2. Replace pass-by-reference with pass-by-value
   - Requires transferring entire object/array data over network
   - Read-only optimization: don't return data if unchanged on server
3. Passing global references
   - Example: file handle to file accessible by client and server via shared file system

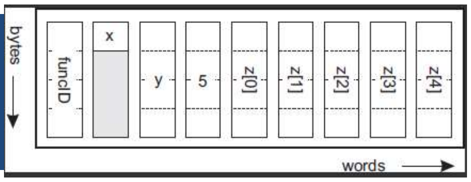


RPC: DEVELOPMENT SUPPORT

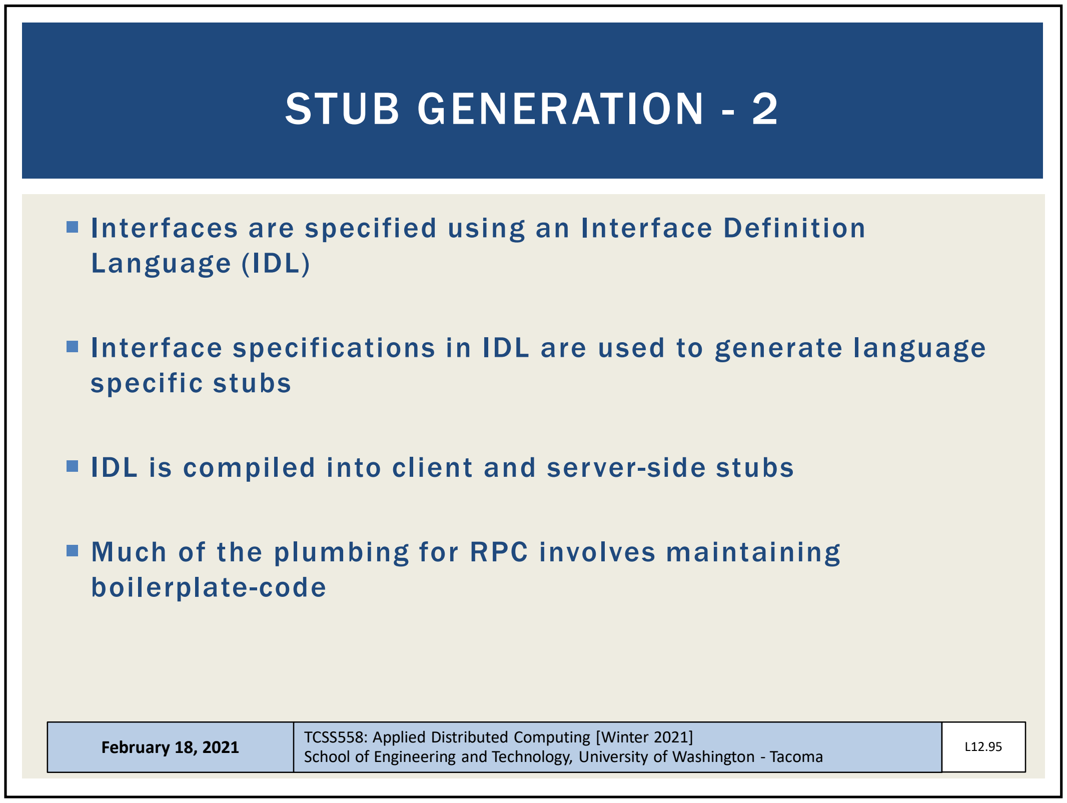
- Let developer specify which routines will be called remotely
  - Automate client/server side stub generation for these routines
- Embed remote procedure call mechanism into the programming language
  - E.g. Java RMI



STUB GENERATION



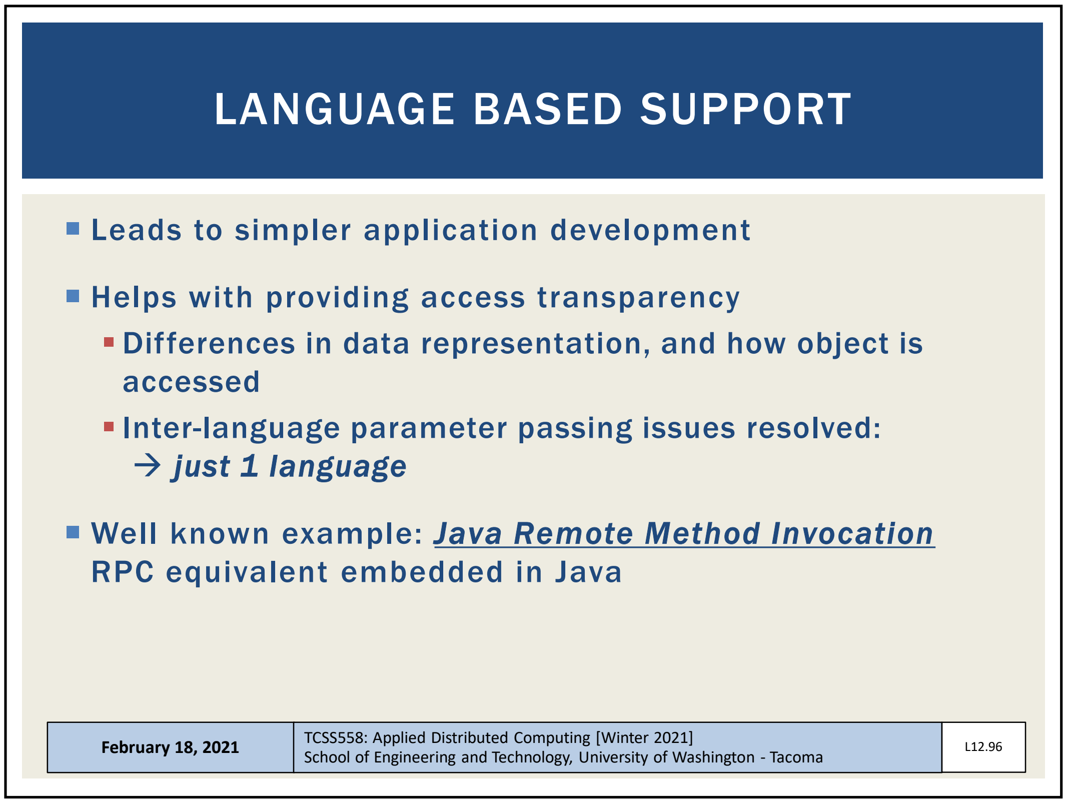
- `void func(char x; float y; int z[5])`
- 1-byte character transmits with 3-padded bytes
- Float sent as whole word (4-bytes)
  - Array as group of words, proceed by word describing length
  - Client stub must package data in specific format
  - Server stub must receive and unpackage in specific format
- Client and server must agree on representation of simple data structures: int, char, floats w/ little endian
- RPC clients/servers: must agree on protocol
  - TCP? UDP?



STUB GENERATION - 2

- Interfaces are specified using an Interface Definition Language (IDL)
- Interface specifications in IDL are used to generate language specific stubs
- IDL is compiled into client and server-side stubs
- Much of the plumbing for RPC involves maintaining boilerplate-code

February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.95



LANGUAGE BASED SUPPORT

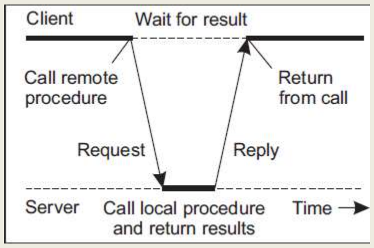
- Leads to simpler application development
- Helps with providing access transparency
  - Differences in data representation, and how object is accessed
  - Inter-language parameter passing issues resolved:
    → ***just 1 language***
- Well known example: ***Java Remote Method Invocation***
  RPC equivalent embedded in Java

February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.96

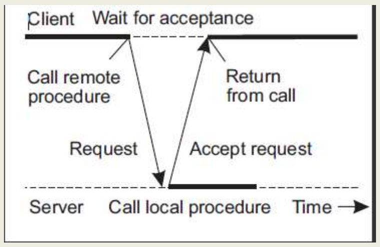
RPC VARIATIONS

- RPC: client typically blocks until reply is returned
- Strict blocking ***unnecessary*** when there is no result
- Asynchronous RPCs
  - When no result, server can immediately send reply

Client/server synchronous RPC



Client/server asynchronous RPC

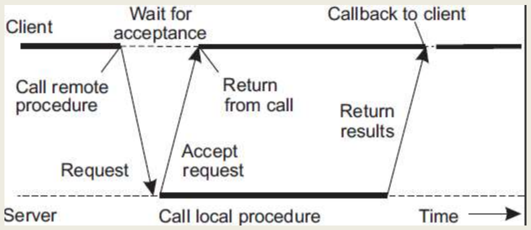


RPC VARIATIONS – 2

- What are tradeoffs for synchronous vs. asynchronous procedure calls?
  - For a local program
  - For a distributed program (system)
- Use cases for asynchronous procedure calls
  - Long running jobs allow client to perform alternate work in background (in parallel)
  - Client may need to make multiple service calls to multiple server backends at the same time…

TYPES OF ASYNCHRONOUS RPC

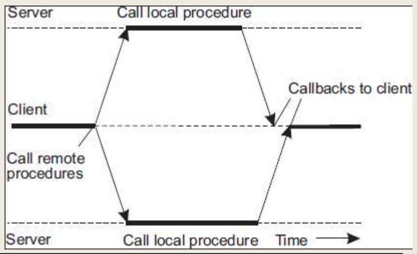
- Deferred synchronous RPC
  - Server performs ***CALLBACK*** to client
  - Client, upon making call, spawns separate thread which blocks and waits for call



- One-way RPCs
  - Client **does not wait** for ***any*** server acknowledgement – it just goes…
- Client polling
  - Client (*using separate thread*) continually polls server for result

MULTICAST RPC

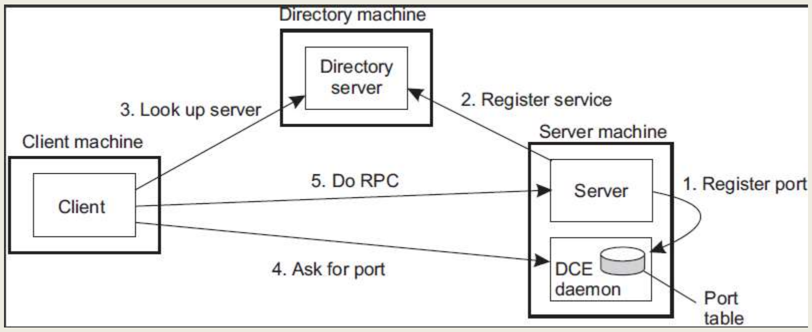
- Send RPC request *simultaneously* to group of servers
- Hide that multiple servers are involved
- Consideration:
  Does the client need all results or just one?
- Use cases:
  - Fault tolerance – wait for just one
  - Replicate execution – verify results, *use first result*
  - Divide and conquer - multiple RPC calls work in parallel on different parts of dataset, client aggregates results



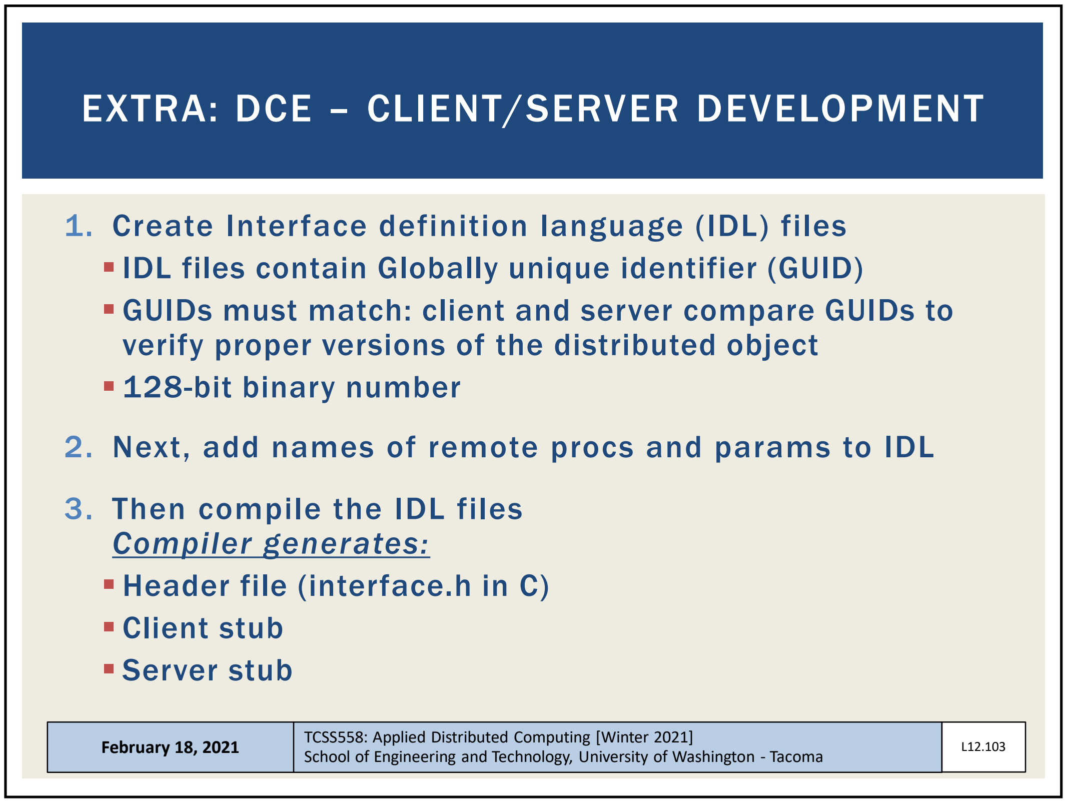
RPC EXAMPLE: DISTRIBUTED COMPUTING ENVIRONMENT (DCE)

- DCE: basis for Microsoft's distributed computing object model (DCOM)
- Used in Samba, ***cross-platform*** file and print sharing via RPC
- Middleware system – provides layer of abstraction between OS and distributed applications
- Designed for Unix, ported to ***all*** major operating systems
- Install DCE middleware on set of heterogeneous machines – distributed applications can then access shared resources to:
  - Mount a windows file system on Linux
  - Share a printer connected to a Windows server
- Uses client/server model
- All communication via RPC
- DCE daemon tracks participating machines, ports

DCE CLIENT-TO-SERVER BINDING

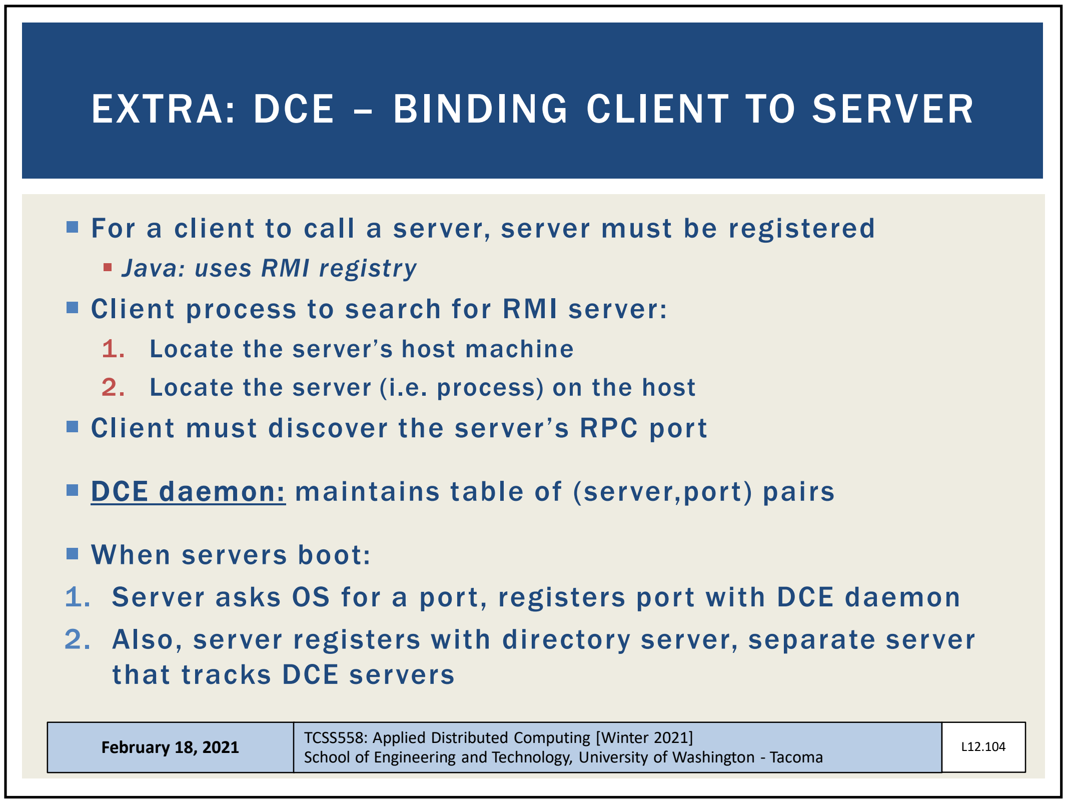


- Server name comes from directory server
- Server port comes from DCE daemon
  - DCE daemon has a well known port # client already knows



EXTRA: DCE – CLIENT/SERVER DEVELOPMENT

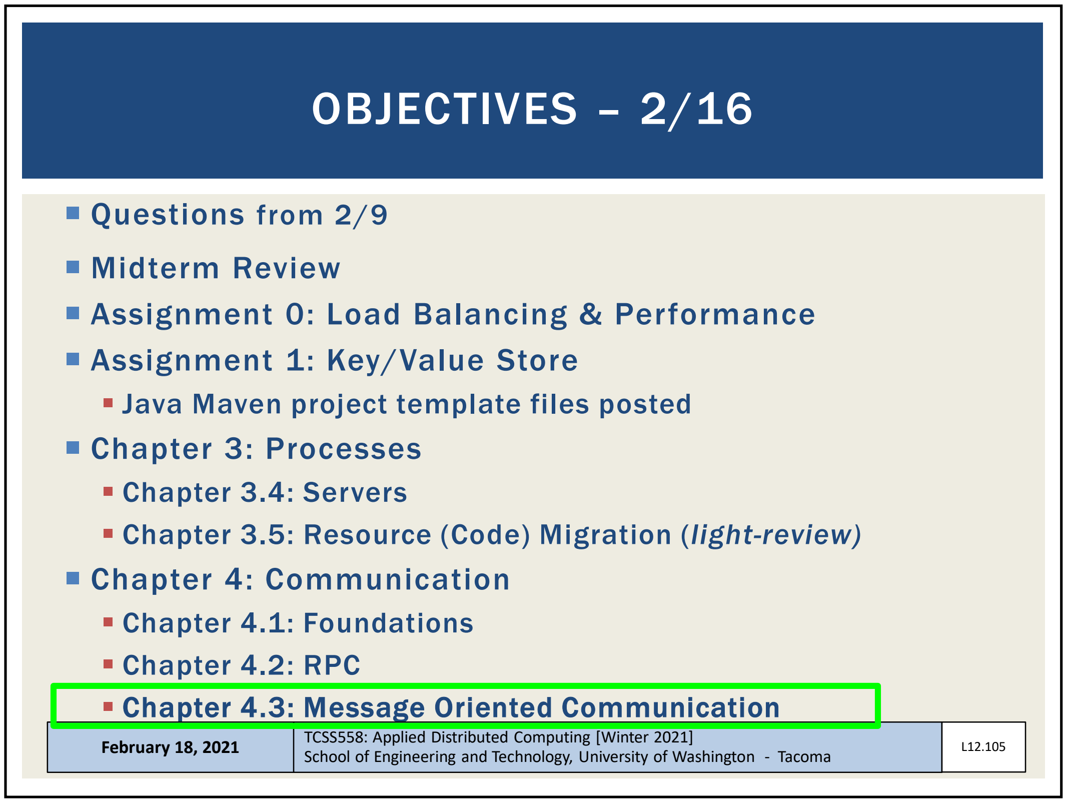
1. Create Interface definition language (IDL) files
   - IDL files contain Globally unique identifier (GUID)
   - GUIDs must match: client and server compare GUIDs to verify proper versions of the distributed object
   - 128-bit binary number
2. Next, add names of remote procs and params to IDL
3. Then compile the IDL files
   Compiler generates:
   - Header file (interface.h in C)
   - Client stub
   - Server stub



EXTRA: DCE – BINDING CLIENT TO SERVER

- For a client to call a server, server must be registered
  - *Java: uses RMI registry*
- Client process to search for RMI server:
  1. Locate the server's host machine
  2. Locate the server (i.e. process) on the host
- Client must discover the server's RPC port
- **DCE daemon:** maintains table of (server,port) pairs
- When servers boot:

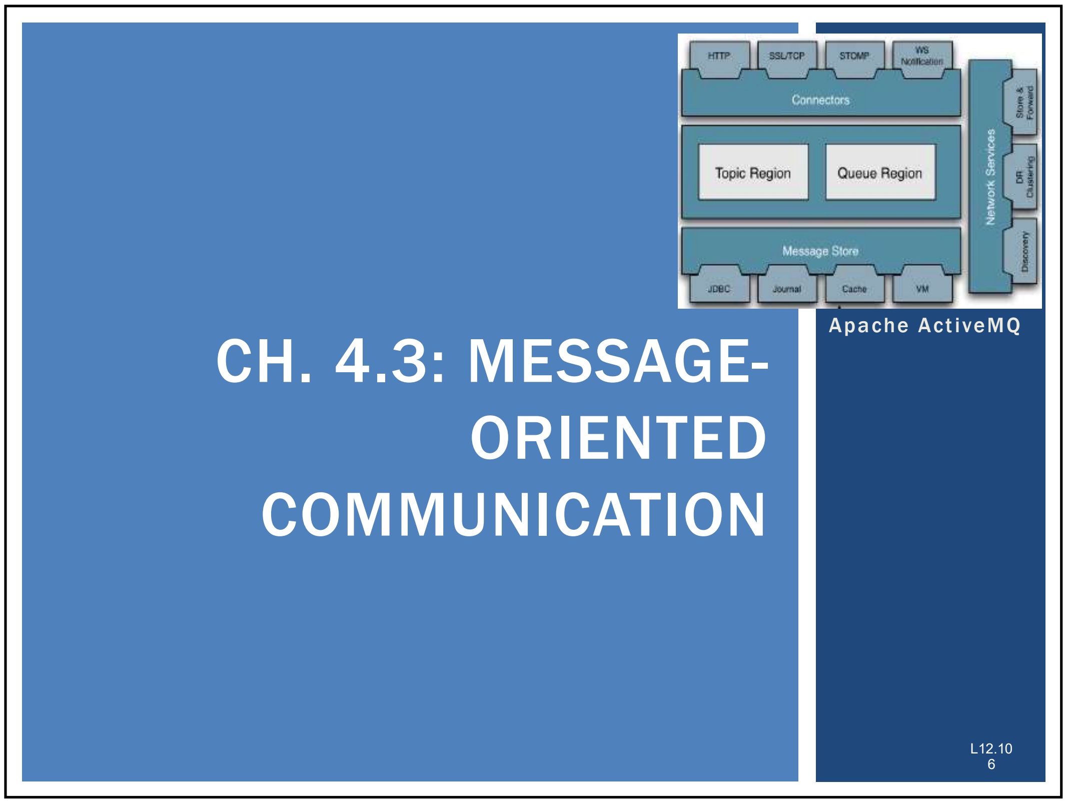
1. Server asks OS for a port, registers port with DCE daemon
2. Also, server registers with directory server, separate server that tracks DCE servers

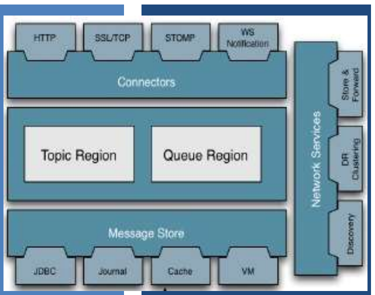


OBJECTIVES – 2/16

- Questions from 2/9
- Midterm Review
- Assignment 0: Load Balancing & Performance
- Assignment 1: Key/Value Store
  - Java Maven project template files posted
- Chapter 3: Processes
  - Chapter 3.4: Servers
  - Chapter 3.5: Resource (Code) Migration (*light-review)*
- Chapter 4: Communication
  - Chapter 4.1: Foundations
  - Chapter 4.2: RPC
  - Chapter 4.3: Message Oriented Communication

February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.105

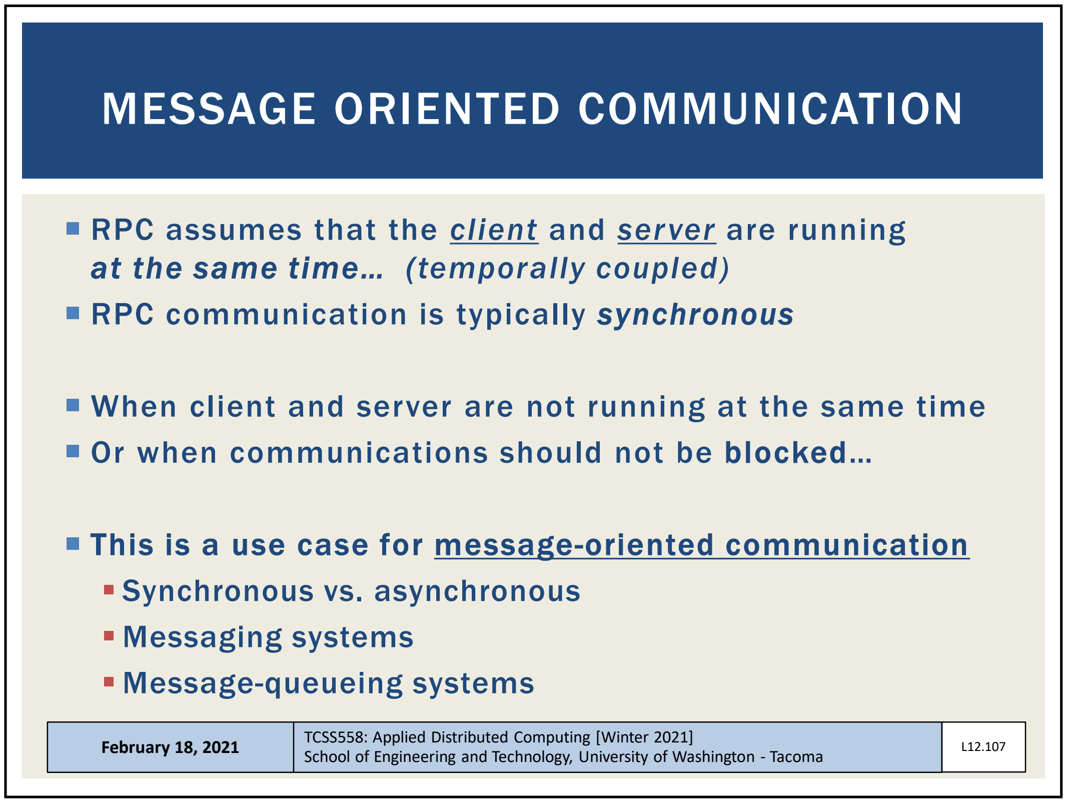




Apache ActiveMQ

CH. 4.3: MESSAGE-ORIENTED COMMUNICATION

L12.106

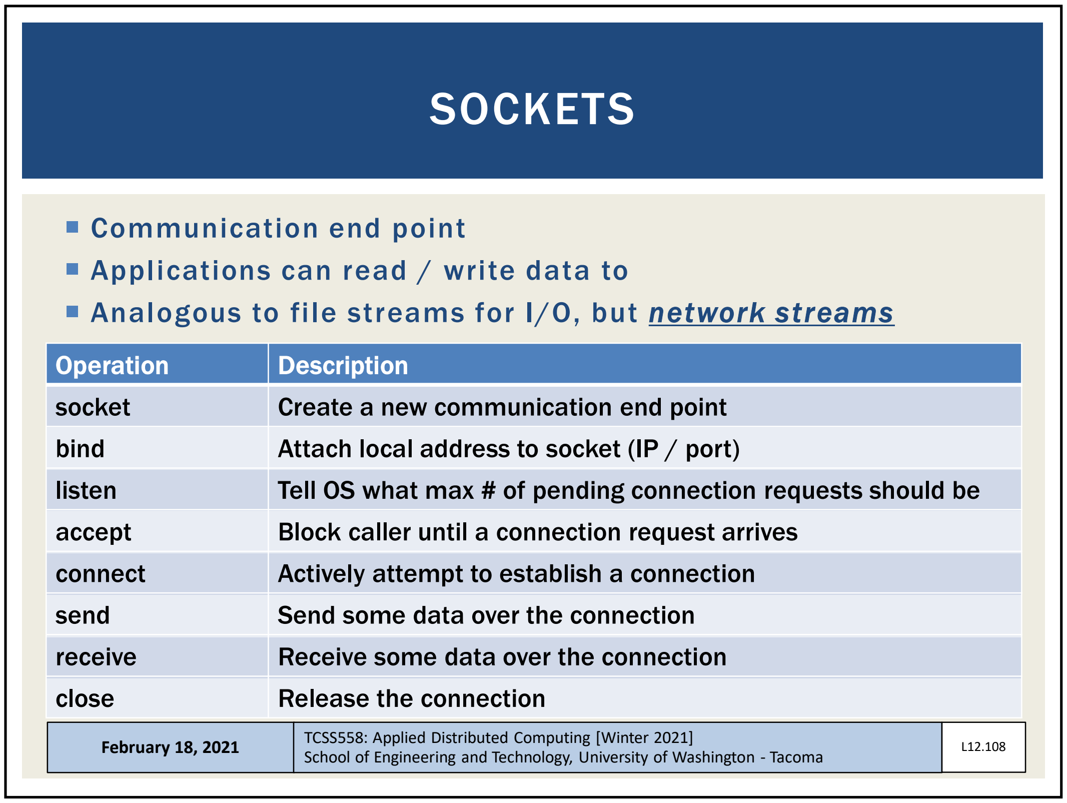


MESSAGE ORIENTED COMMUNICATION

- RPC assumes that the *client* and *server* are running ***at the same time… (temporally coupled)***
- RPC communication is typically ***synchronous***

- When client and server are not running at the same time
- Or when communications should not be **blocked**…

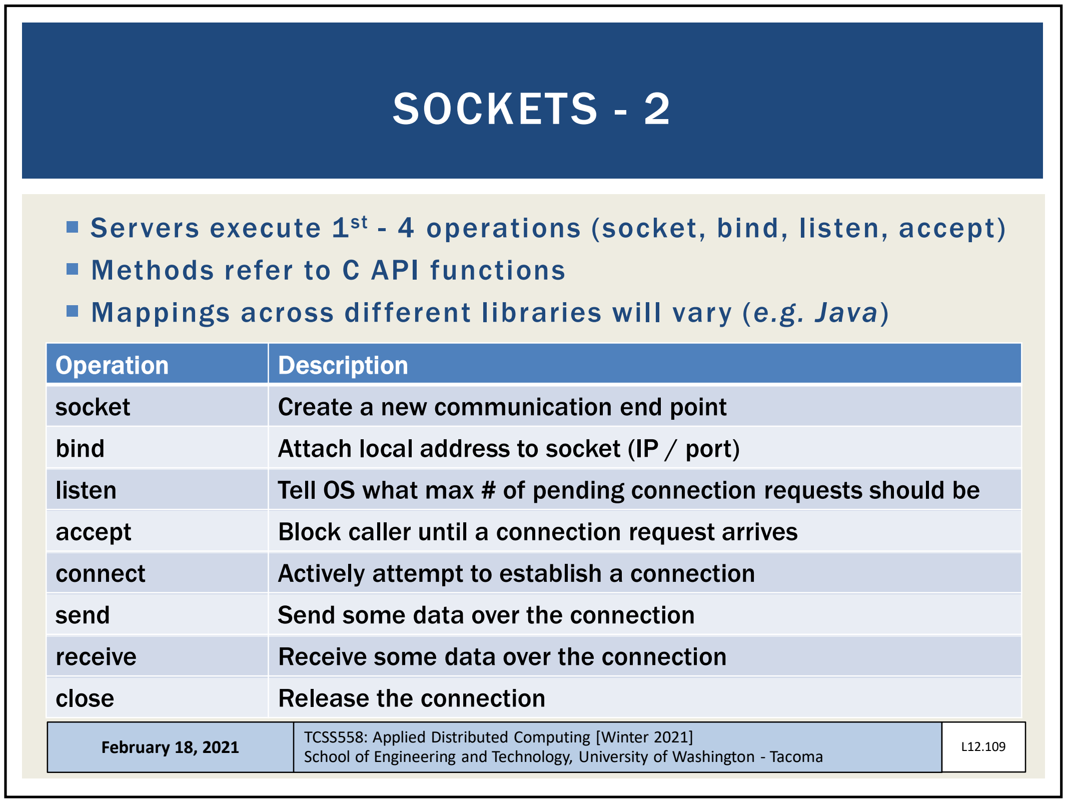
- **This is a use case for message-oriented communication**
  - Synchronous vs. asynchronous
  - Messaging systems
  - Message-queueing systems



SOCKETS

- Communication end point
- Applications can read / write data to
- Analogous to file streams for I/O, but ***network streams***

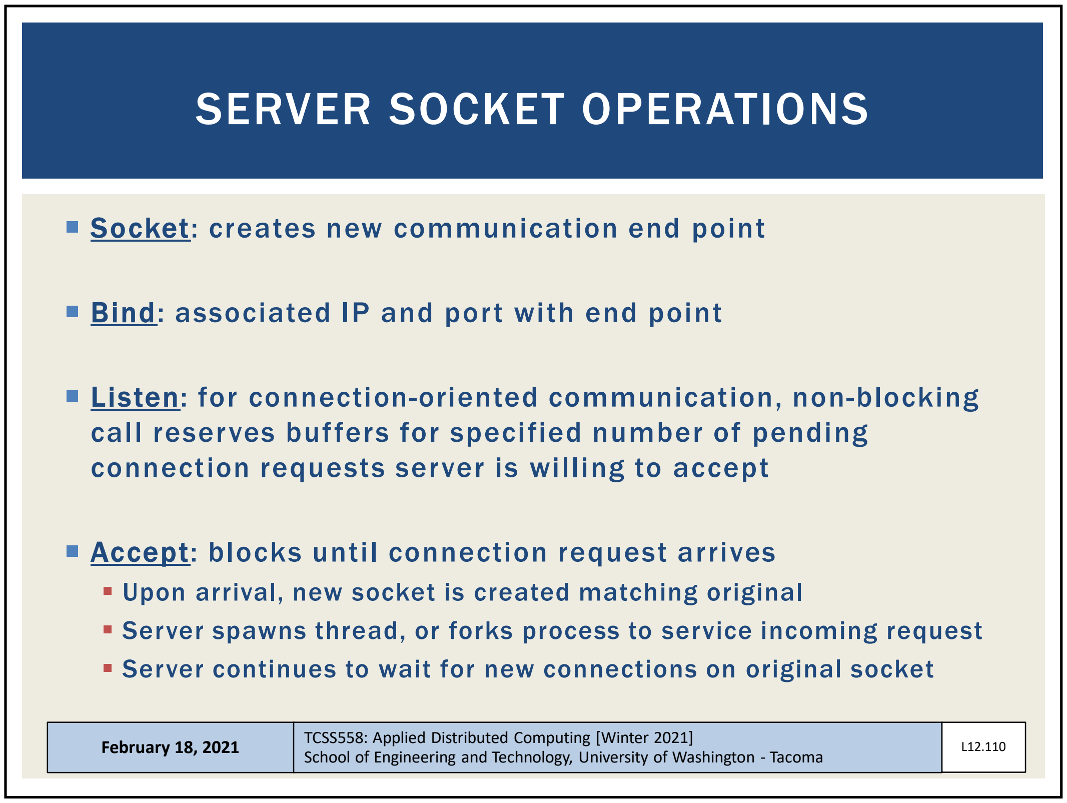
Operation	Description
socket	Create a new communication end point
bind	Attach local address to socket (IP / port)
listen	Tell OS what max # of pending connection requests should be
accept	Block caller until a connection request arrives
connect	Actively attempt to establish a connection
send	Send some data over the connection
receive	Receive some data over the connection
close	Release the connection



SOCKETS - 2

- Servers execute 1st - 4 operations (socket, bind, listen, accept)
- Methods refer to C API functions
- Mappings across different libraries will vary (*e.g. Java*)

Operation	Description
socket	Create a new communication end point
bind	Attach local address to socket (IP / port)
listen	Tell OS what max # of pending connection requests should be
accept	Block caller until a connection request arrives
connect	Actively attempt to establish a connection
send	Send some data over the connection
receive	Receive some data over the connection
close	Release the connection

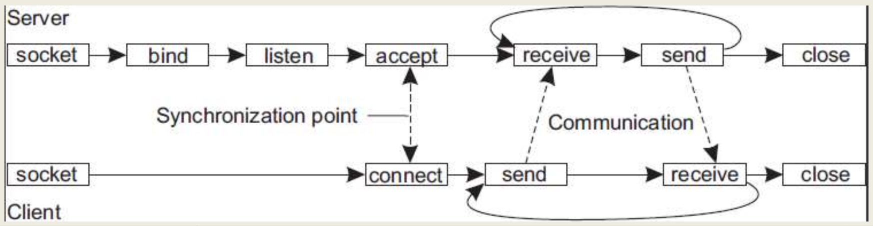


SERVER SOCKET OPERATIONS

- Socket: creates new communication end point
- Bind: associated IP and port with end point
- Listen: for connection-oriented communication, non-blocking call reserves buffers for specified number of pending connection requests server is willing to accept
- Accept: blocks until connection request arrives
  - Upon arrival, new socket is created matching original
  - Server spawns thread, or forks process to service incoming request
  - Server continues to wait for new connections on original socket

CLIENT SOCKET OPERATIONS

- **Socket**: Creates socket client uses for communication
- **Connect**: Server transport-level address provided, client blocks until connection established
- **Send**: Supports sending data (to: server/client)
- **Receive**: Supports receiving data (from: server/client)
- **Close**: Closes communication channel
  - Analogous to closing a file stream

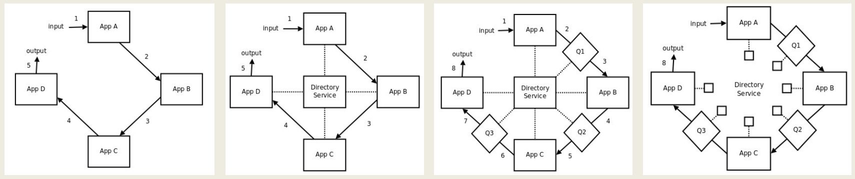


SOCKET COMMUNICATION

- Sockets provide primitives for implementing your own TCP/UDP communication protocols
- Directly using sockets for transient (non-persisted) messaging is very basic, can be brittle
  - Easy to make mistakes…
- Any extra communication facilities must be implemented by the application developer
- More advanced approaches are desirable
  - E.g. frameworks with support common desirable functionality

ZEROMQ – SOCKET LIBRARY

- (0MQ) High performance intelligent <u>socket library</u>
- *<u>zero broker, zero latency, zero admin, zero cost, zero waste</u>*
- Provides a message queue
- ***Builds upon*** functionality of <u>traditional sockets</u>
- Implementation in C++
  - 30+ language bindings provided
- Enables support for various messaging patterns
- Can support brokered (centralized) and broker-less topologies

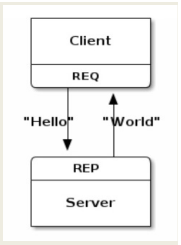


ZEROMQ – 2

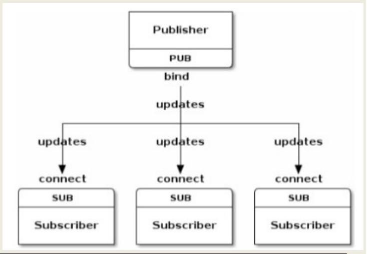
- ZeroMQ is <u>TCP-connection-oriented communication</u>
- Provides socket-like primitives with more functionality
  - Basic socket operations ***abstracted*** away
  - Supports many-to-one, one-to-one, and one-to-many connections
  - ***Multicast*** connections (one-to-many – single server socket simultaneously "connects" to multiple clients)
- Asynchronous messaging
- Supports pairing sockets to support communication patterns

ZEROMQ - PATTERNS

- **Request-reply pattern**
  - Traditional client-server communication (e.g. RPC)
  - Client: request socket (**REQ**)
  - Server: reply socket (**REP**)

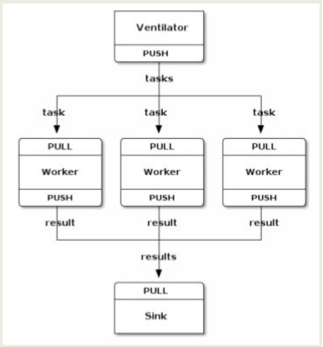


- **Publish-subscribe pattern**
  - Clients **subscribe** to messages **published** by servers
  - As in event-based coordination (Ch. 1)
  - Supports multicasting messages from server to multiple
  - Client: subscribe socket (**SUB**)
  - Server: publish socket (**PUB**)



ZEROMQ – PATTERNS - 2

- **Pipeline pattern (FIFO-queue)**
  - Analogous to a producer/consumer bounded buffer
  - Producing processes generate results, push to pipe
  - Consuming processes consume results, pull from pipe
  - Producers: push socket (**PUSH** socket)
  - Consumers: pull socket (**PULL** socket)
  - Push- distributes messages to all pull clients evenly
  - Consumers pull results from pipe and push results downstream

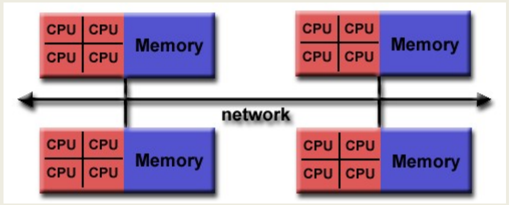


QUEUEING ALTERNATIVES

- Cloud services
  - Amazon Simple Queueing Service (SQS)
  - Azure service bus
- Open source frameworks
  - Nanomsg
  - ZeroMQ

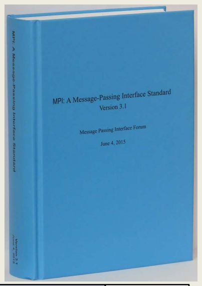
MESSAGE PASSING INTERFACE (MPI)

- MPI introduced – version 1.0 March 1994
- Message passing API for parallel programming: ***supercomputers***
- Communication protocol for parallel programming for: Supercomputers, High Performance Computing (HPC) clusters
- Point-to-point and collective communication
- Goals: high performance, scalability, portability
- Most implementations in C, C++, Fortran



MOTIVATIONS FOR MPI

- Motivation: sockets insufficient for interprocess communication on large scale HPC compute clusters and super computers
  - Sockets at the wrong level of abstraction
  - Sockets designed to communicate over the network using general purpose TCP/IP stacks
  - Not designed for proprietary protocols
  - Not designed for high-speed interconnection networks used by supercomputers, HPC-clusters, etc.
  - Better buffering and synchronization needed

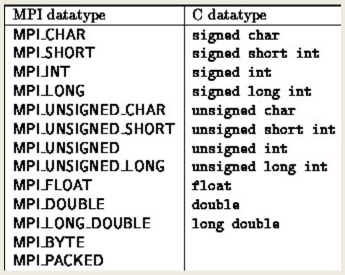


MOTIVATIONS FOR MPI - 2

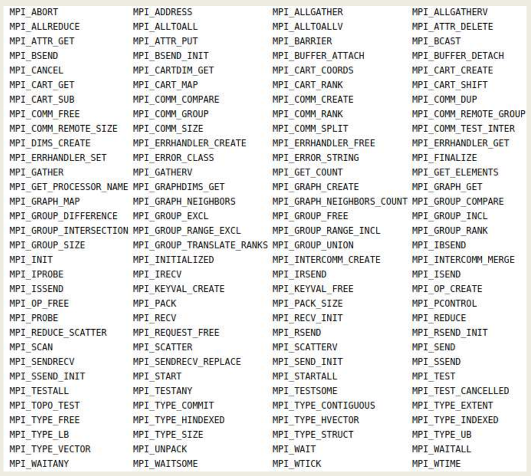
- Supercomputers had proprietary communication libraries
  - Offer a wealth of efficient communication operations
- All libraries mutually incompatible
- Led to significant portability problems developing parallel code that could migrate across supercomputers
- Led to development of MPI
  - To support transient (non-persistent) communication for parallel programming

MPI FUNCTIONS / DATATYPES

- Very large library, v1.0 (1994) 128 functions
- Version 3 (2015) 440+
- MPI data types:
- Provide common mappings



MPI datatype	C datatype
MPI_CHAR	signed char
MPI_SHORT	signed short int
MPI_INT	signed int
MPI_LONG	signed long int
MPI_UNSIGNED_CHAR	unsigned char
MPI_UNSIGNED_SHORT	unsigned short int
MPI_UNSIGNED	unsigned int
MPI_UNSIGNED_LONG	unsigned long int
MPI_FLOAT	float
MPI_DOUBLE	double
MPI_LONG_DOUBLE	long double
MPI_BYTE	
MPI_PACKED	

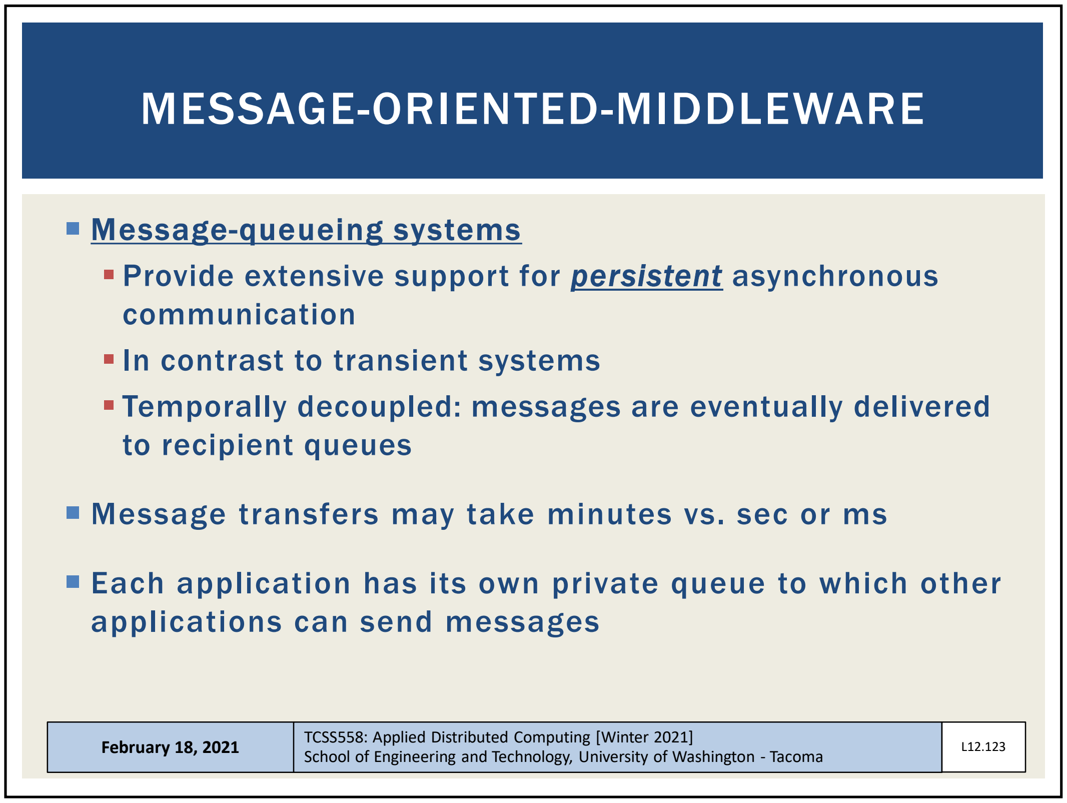


MPI_ABORT	MPI_ADDRESS	MPI_ALLGATHER	MPI_ALLGATHERV
MPI_ALLREDUCE	MPI_ALLTOALL	MPI_ALLTOALLV	MPI_ATTR_DELETE
MPI_ATTR_GET	MPI_ATTR_PUT	MPI_BARRIER	MPI_BCAST
MPI_BSEND	MPI_BSEND_INIT	MPI_BUFFER_ATTACH	MPI_BUFFER_DETACH
MPI_CANCEL	MPI_CARTDIM_GET	MPI_CART_COORDS	MPI_CART_CREATE
MPI_CART_GET	MPI_CART_MAP	MPI_CART_RANK	MPI_CART_SHIFT
MPI_CART_SUB	MPI_COMM_COMPARE	MPI_COMM_CREATE	MPI_COMM_DUP
MPI_COMM_FREE	MPI_COMM_GROUP	MPI_COMM_RANK	MPI_COMM_REMOTE_GROUP
MPI_COMM_REMOTE_SIZE	MPI_COMM_SIZE	MPI_COMM_SPLIT	MPI_COMM_TEST_INTER
MPI_DIMS_CREATE	MPI_ERRHANDLER_CREATE	MPI_ERRHANDLER_FREE	MPI_ERRHANDLER_GET
MPI_ERRHANDLER_SET	MPI_ERROR_CLASS	MPI_ERROR_STRING	MPI_FINALIZE
MPI_GATHER	MPI_GATHERV	MPI_GET_COUNT	MPI_GET_ELEMENTS
MPI_GET_PROCESSOR_NAME	MPI_GRAPHDIMS_GET	MPI_GRAPH_CREATE	MPI_GRAPH_GET
MPI_GRAPH_MAP	MPI_GRAPH_NEIGHBORS	MPI_GRAPH_NEIGHBORS_COUNT	MPI_GROUP_COMPARE
MPI_GROUP_DIFFERENCE	MPI_GROUP_EXCL	MPI_GROUP_FREE	MPI_GROUP_INCL
MPI_GROUP_INTERSECTION	MPI_GROUP_RANGE_EXCL	MPI_GROUP_RANGE_INCL	MPI_GROUP_RANK
MPI_GROUP_SIZE	MPI_GROUP_TRANSLATE_RANKS	MPI_GROUP_UNION	MPI_IBSEND
MPI_INIT	MPI_INITIALIZED	MPI_INTERCOMM_CREATE	MPI_INTERCOMM_MERGE
MPI_IPROBE	MPI_IRECV	MPI_IRSEND	MPI_ISEND
MPI_ISSEND	MPI_KEYVAL_CREATE	MPI_KEYVAL_FREE	MPI_OP_CREATE
MPI_OP_FREE	MPI_PACK	MPI_PACK_SIZE	MPI_PCONTROL
MPI_PROBE	MPI_RECV	MPI_RECV_INIT	MPI_REDUCE
MPI_REDUCE_SCATTER	MPI_REQUEST_FREE	MPI_RSEND	MPI_RSEND_INIT
MPI_SCAN	MPI_SCATTER	MPI_SCATTERV	MPI_SEND
MPI_SENDRECV	MPI_SENDRECV_REPLACE	MPI_SEND_INIT	MPI_SSEND
MPI_SSEND_INIT	MPI_START	MPI_STARTALL	MPI_TEST
MPI_TESTALL	MPI_TESTANY	MPI_TESTSOME	MPI_TEST_CANCELLED
MPI_TOPO_TEST	MPI_TYPE_COMMIT	MPI_TYPE_CONTIGUOUS	MPI_TYPE_EXTENT
MPI_TYPE_FREE	MPI_TYPE_HINDEXED	MPI_TYPE_HVECTOR	MPI_TYPE_INDEXED
MPI_TYPE_LB	MPI_TYPE_SIZE	MPI_TYPE_STRUCT	MPI_TYPE_UB
MPI_TYPE_VECTOR	MPI_UNPACK	MPI_WAIT	MPI_WAITALL
MPI_WAITANY	MPI_WAITSOME	MPI_WTICK	MPI_WTIME

COMMON MPI FUNCTIONS

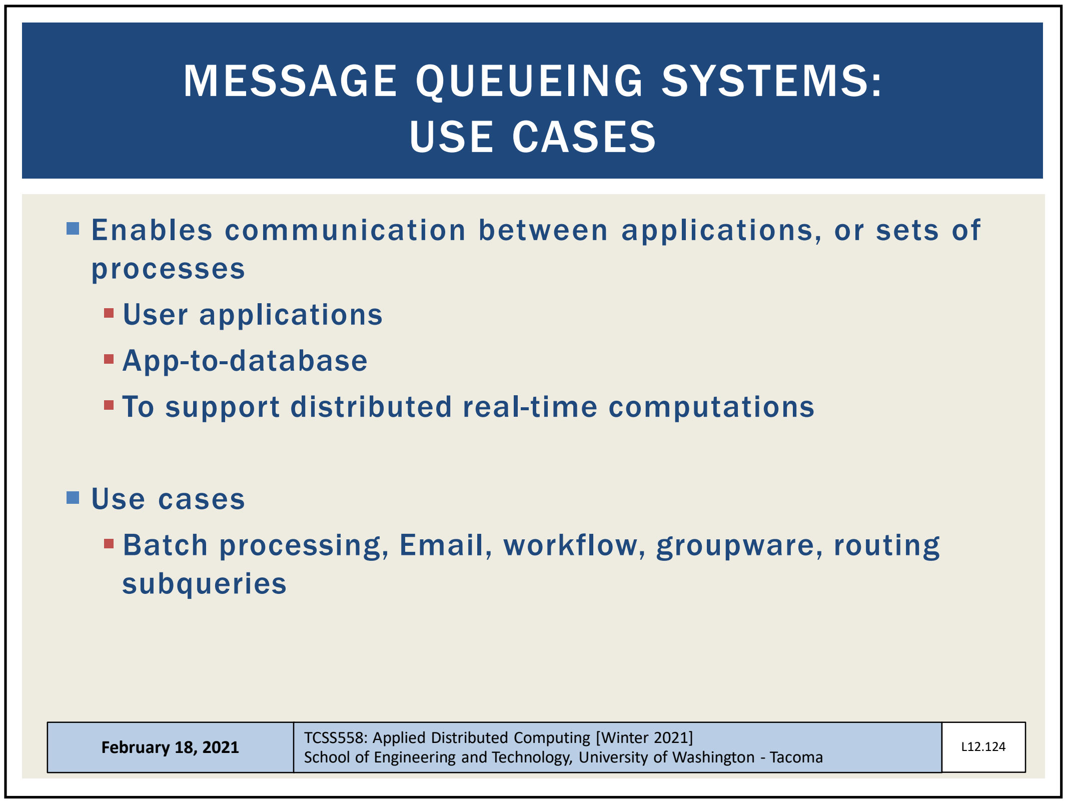
- MPI - no recovery for process crashes, network partitions
- Communication among grouped processes: `(groupID, processID)`
- IDs used to route messages in place of IP addresses

Operation	Description
MPI_bsend	Append outgoing message to a local send buffer
MPI_send	Send message, wait until copied to local/remote buffer
MPI_ssend	Send message, wat until transmission starts
MPI_sendrecv	Send message, wait for reply
MPI_isend	Pass reference to outgoing message and continue
MPI_issend	Pass reference to outgoing messages, wait until receipt start
MPI_recv	Receive a message, block if there is none
MPI_irecv	Check for incoming message, **do not block!**



MESSAGE-ORIENTED-MIDDLEWARE

- <u>**Message-queueing systems**</u>
  - Provide extensive support for ***<u>persistent</u>*** asynchronous communication
  - In contrast to transient systems
  - Temporally decoupled: messages are eventually delivered to recipient queues
- Message transfers may take minutes vs. sec or ms
- Each application has its own private queue to which other applications can send messages

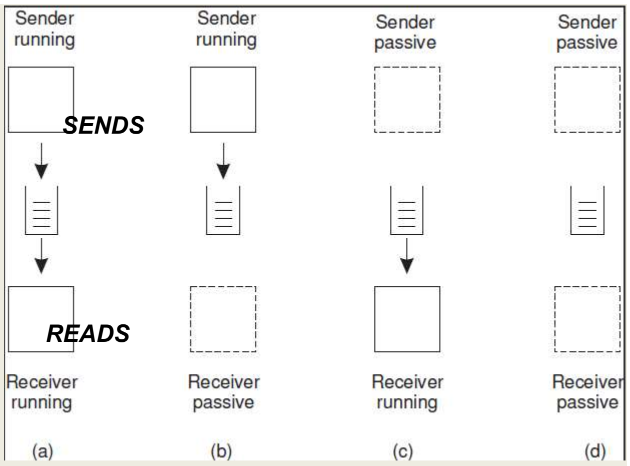


MESSAGE QUEUEING SYSTEMS: USE CASES

- Enables communication between applications, or sets of processes
  - User applications
  - App-to-database
  - To support distributed real-time computations
- Use cases
  - Batch processing, Email, workflow, groupware, routing subqueries

MESSAGE QUEUEING SYSTEMS

- Scenarios:
  (a) Sender/receiver both running
  (b) Sender running, receiver offline
  (c) Sender offline, receiver running
  (d) Sender/receiver both offline



- Queue persists msgs, and attempts to send them but no one may be available to receive them…

MESSAGE QUEUEING SYSTEMS - 2

- **Key:** Truly persistent messaging
- Message queueing systems can persist messages for awhile and senders and receivers can be offline
- **Messages**
- Contain *any* data, may have size limit
- Are properly addressed, to a destination queue
- **Basic Inteface**
- PUT: called by sender to append msg to specified queue
- GET: blocking call to remove oldest msg from specified queue
  - Blocked if queue is empty
- POLL: Non-blocking, gets msg from specified queue

MESSAGE QUEUEING SYSTEMS ARCHITECTURE

- **Basic interface cont'd**
- NOTIFY: install a callback function, for when msg is placed into a queue. Notifies receivers

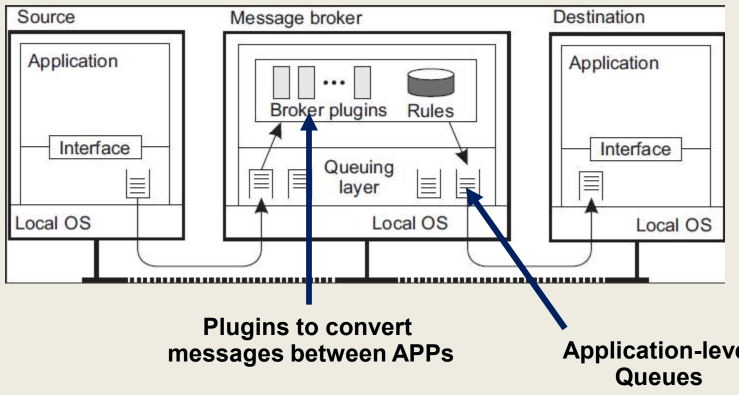
- **Queue managers**: manage individual message queues as a separate process/library
- Applications get/put messages only from ***local*** queues
- Queue manager and apps share local network
- **ISSUES:**
- How should we reference the destination queue?
- How should names be resolved (looked-up)?
  - Contact address (host, port) pairs
  - Local look-up tables can be stored at each queue manager

MESSAGE QUEUEING SYSTEMS ARCHITECTURE - 2

- **ISSUES**:
- How do we route traffic between queue managers?
  - How are name-to-address mappings efficiently kept?
  - Each queue manager should be known to all others

- **Message brokers**
- Handle message conversion among different users/formats
- Addresses cases when senders and receivers don't speak the same protocol (language)
- Need arises for message protocol converters
  - "Reformatter" of messages
- Act as application-level gateway

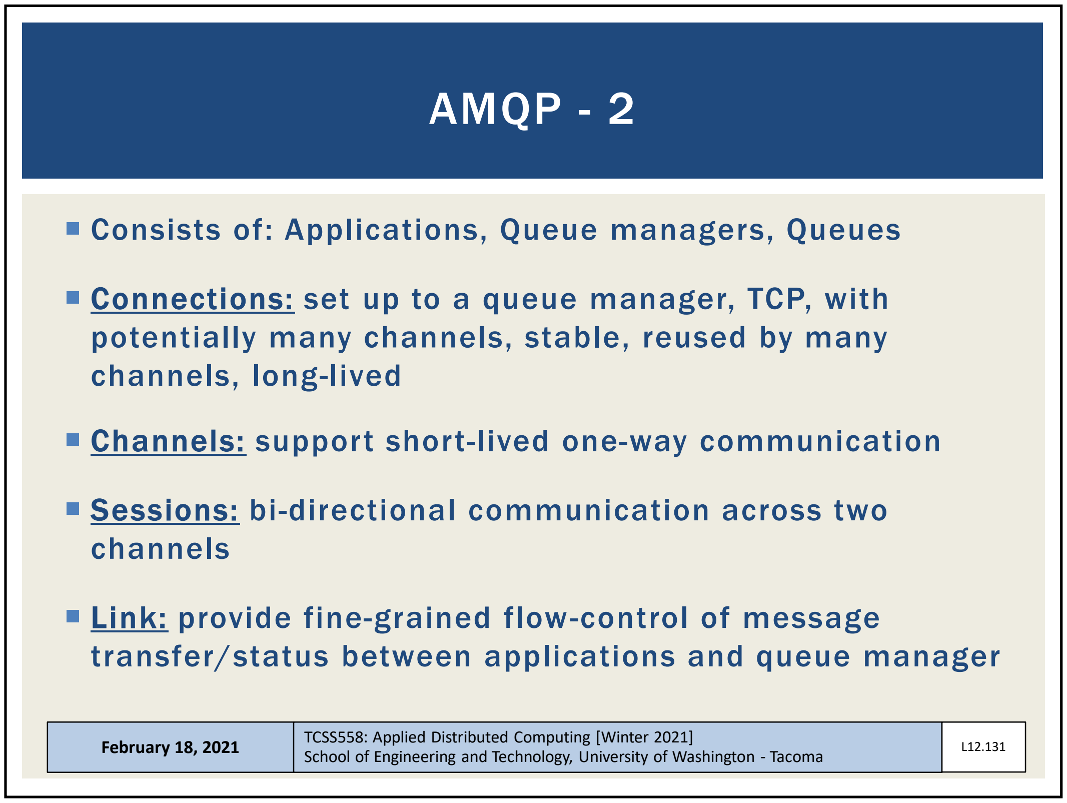
MESSAGE BROKER ORGANIZATION



AMQP PROTOCOL

- Message-queueing systems initially developed to enable legacy applications to interoperate
- Decouple inter-application communication to "open" messaging-middleware
- Many are proprietary solutions, ***so not very open***
- e.g. Microsoft Message Queueing service, Windows NT 1997

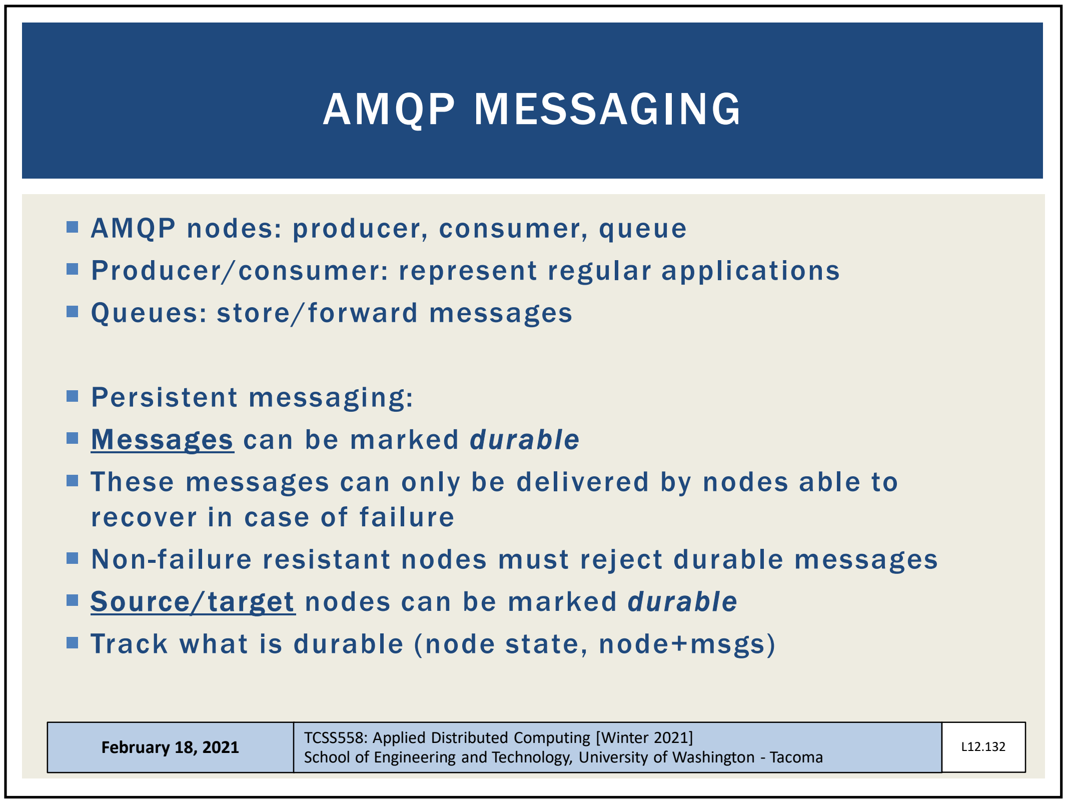
- Advanced message queueing protocol (AMQP), 2006
- Address openness/interoperability of proprietary solutions
- Open wire protocol for messaging with powerful routing capabilities
- Help *abstract* messaging and application interoperability by means of a generic open protocol
- Suffer from incompatibility among protocol versions
- pre-1.0, 1.0+



AMQP - 2

- Consists of: Applications, Queue managers, Queues
- **Connections:** set up to a queue manager, TCP, with potentially many channels, stable, reused by many channels, long-lived
- **Channels:** support short-lived one-way communication
- **Sessions:** bi-directional communication across two channels
- **Link:** provide fine-grained flow-control of message transfer/status between applications and queue manager

February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.131

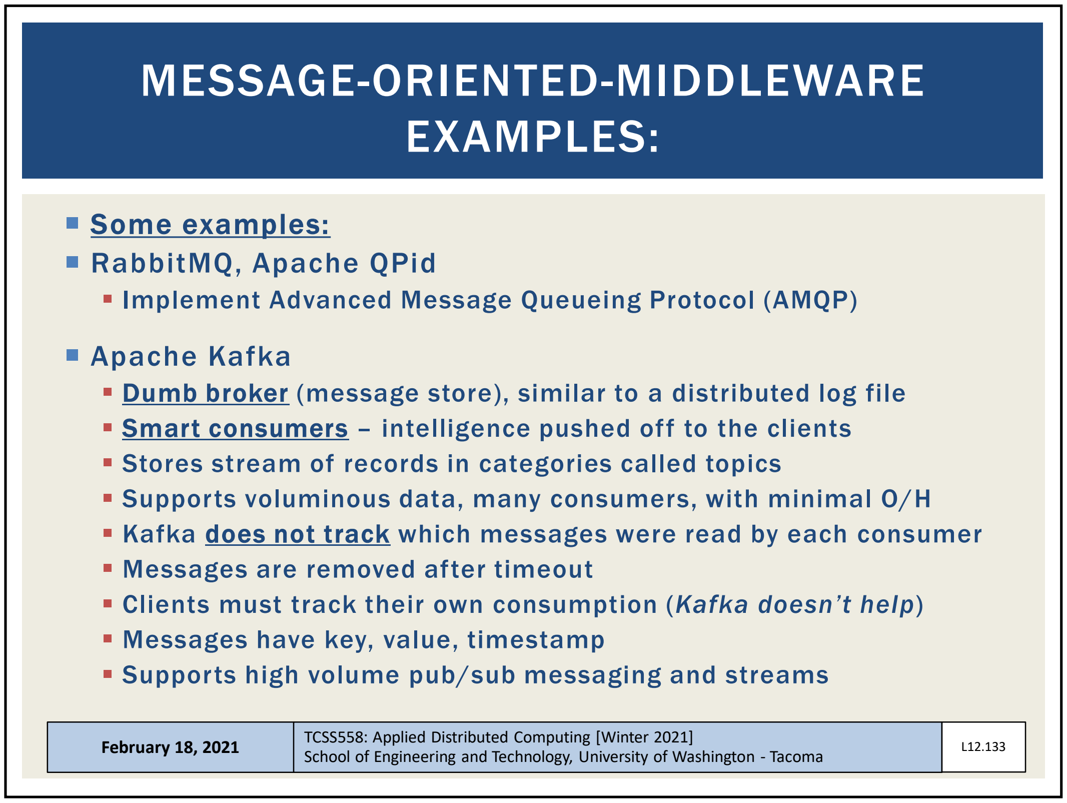


AMQP MESSAGING

- AMQP nodes: producer, consumer, queue
- Producer/consumer: represent regular applications
- Queues: store/forward messages

- Persistent messaging:
- **Messages** can be marked ***durable***
- These messages can only be delivered by nodes able to recover in case of failure
- Non-failure resistant nodes must reject durable messages
- **Source/target** nodes can be marked ***durable***
- Track what is durable (node state, node+msgs)

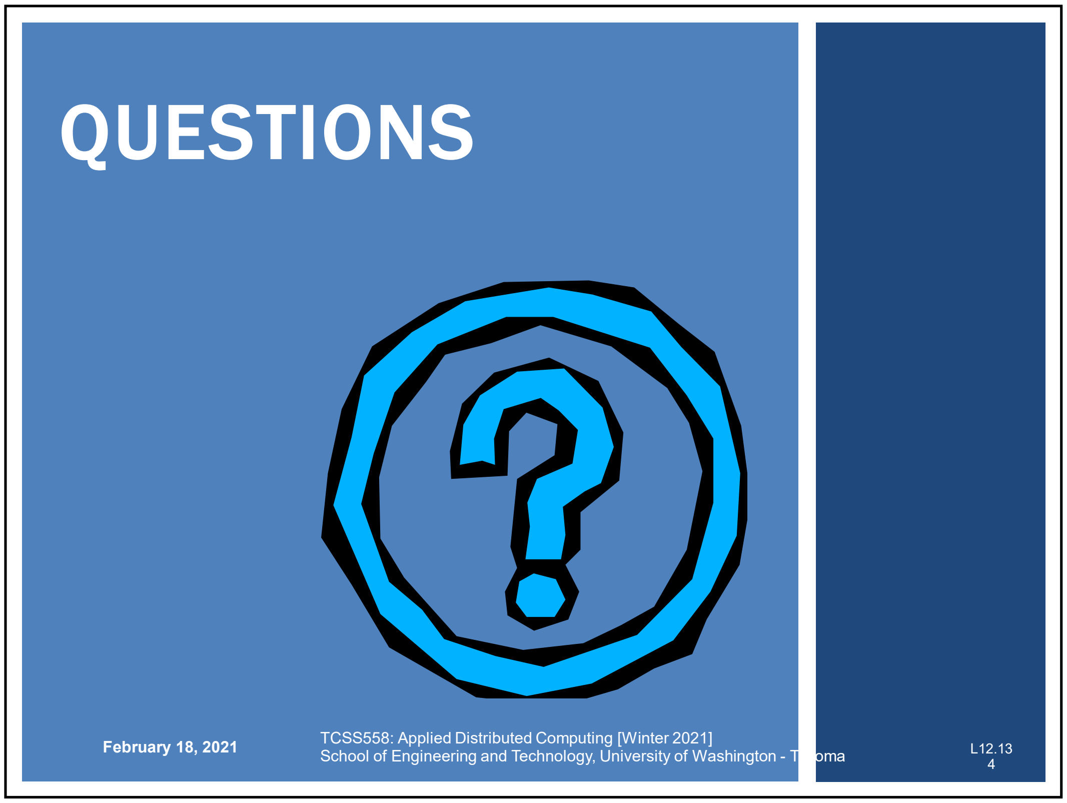
February 18, 2021	TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma	L12.132



MESSAGE-ORIENTED-MIDDLEWARE EXAMPLES:

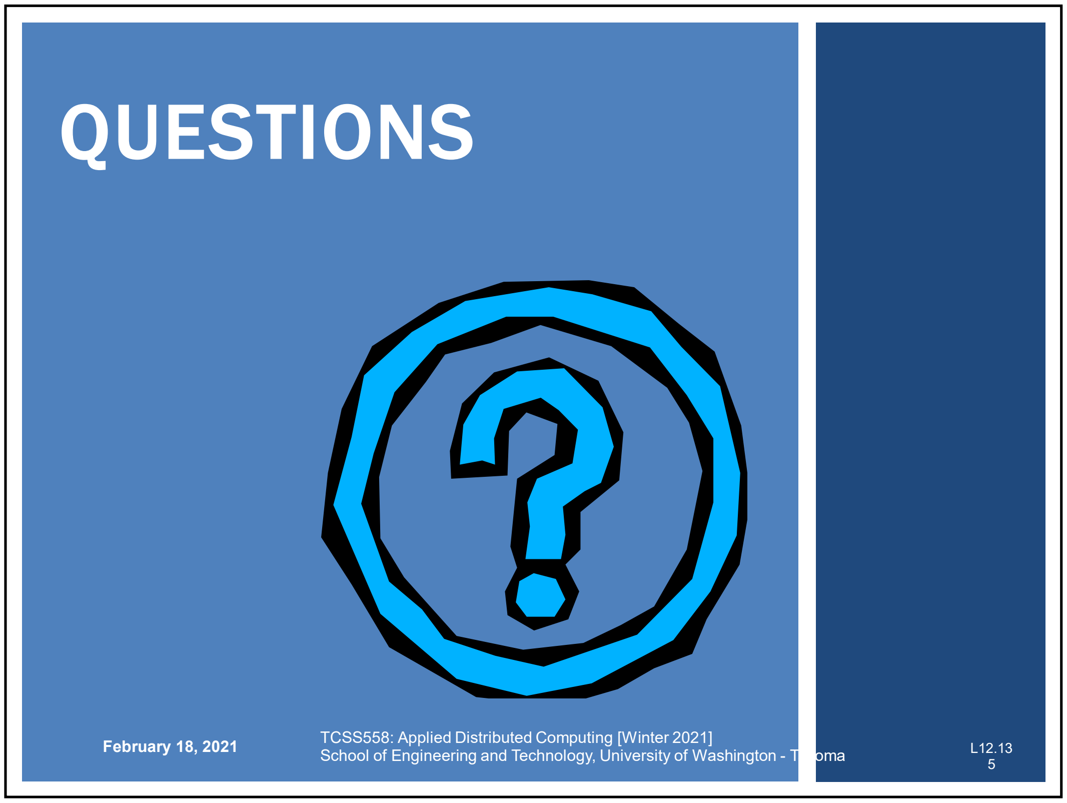
- **Some examples:**
- RabbitMQ, Apache QPid
  - Implement Advanced Message Queueing Protocol (AMQP)
- Apache Kafka
  - **Dumb broker** (message store), similar to a distributed log file
  - **Smart consumers** – intelligence pushed off to the clients
  - Stores stream of records in categories called topics
  - Supports voluminous data, many consumers, with minimal O/H
  - Kafka **does not track** which messages were read by each consumer
  - Messages are removed after timeout
  - Clients must track their own consumption (*Kafka doesn't help*)
  - Messages have key, value, timestamp
  - Supports high volume pub/sub messaging and streams

February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - Tacoma | L12.133

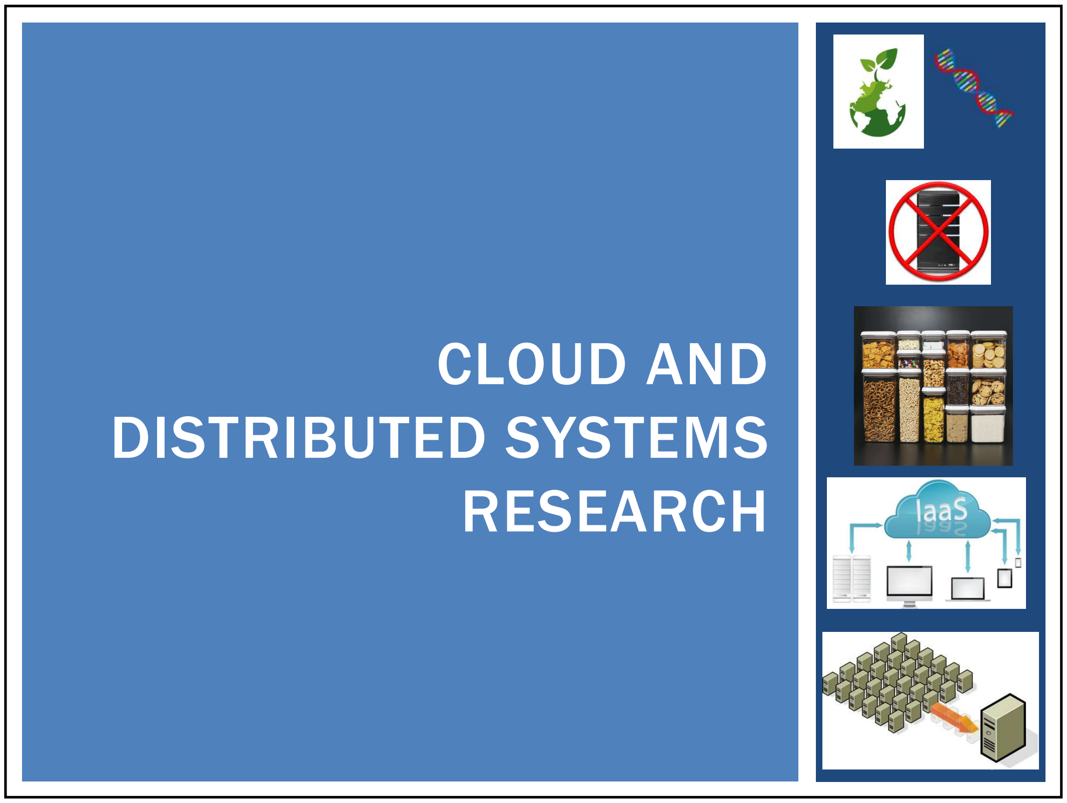


QUESTIONS

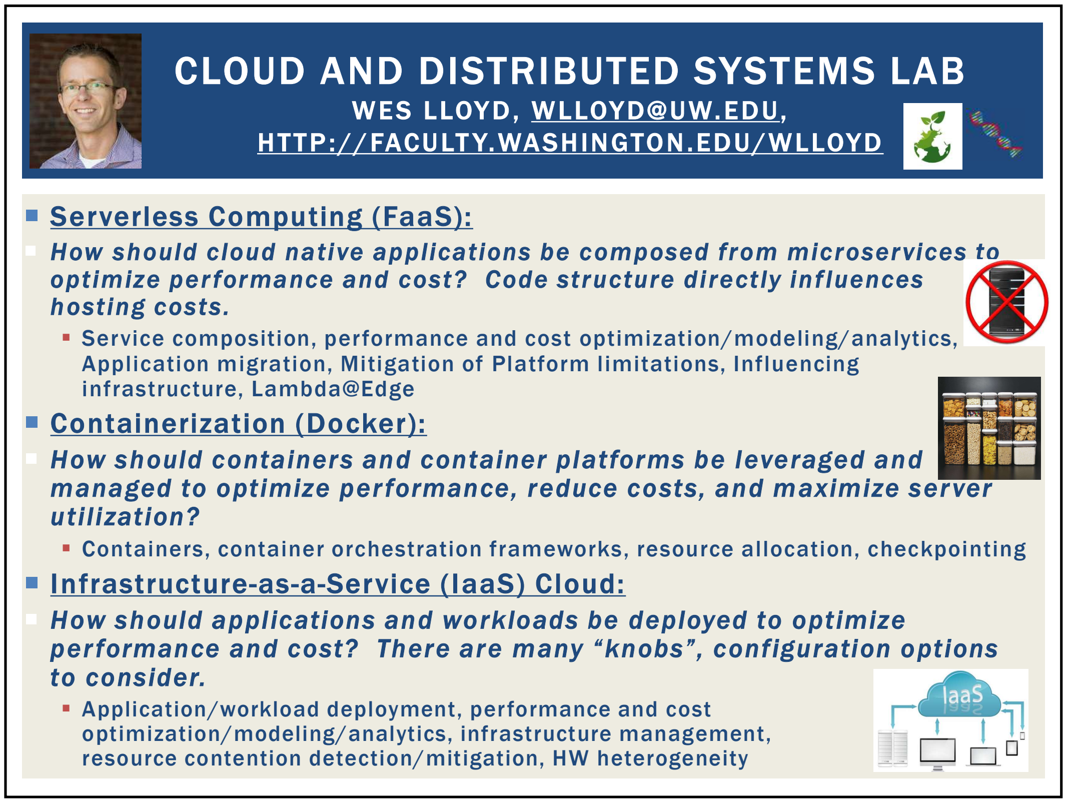
February 18, 2021 | TCSS558: Applied Distributed Computing [Winter 2021] School of Engineering and Technology, University of Washington - T oma | L12.13 4



QUESTIONS



CLOUD AND DISTRIBUTED SYSTEMS RESEARCH



CLOUD AND DISTRIBUTED SYSTEMS LAB

WES LLOYD, WLLOYD@UW.EDU,
HTTP://FACULTY.WASHINGTON.EDU/WLLOYD

- **Serverless Computing (FaaS):**
- ***How should cloud native applications be composed from microservices to optimize performance and cost? Code structure directly influences hosting costs.***
  - Service composition, performance and cost optimization/modeling/analytics, Application migration, Mitigation of Platform limitations, Influencing infrastructure, Lambda@Edge
- **Containerization (Docker):**
- ***How should containers and container platforms be leveraged and managed to optimize performance, reduce costs, and maximize server utilization?***
  - Containers, container orchestration frameworks, resource allocation, checkpointing
- **Infrastructure-as-a-Service (IaaS) Cloud:**
- ***How should applications and workloads be deployed to optimize performance and cost? There are many "knobs", configuration options to consider.***
  - Application/workload deployment, performance and cost optimization/modeling/analytics, infrastructure management, resource contention detection/mitigation, HW heterogeneity

TCSS 558

OFFICE HOURS

PLEASE SAY HELLO

HAVE STEPPED OUT

WILL RETURN SHORTLY